

# *LEAD SCORE CASE STUDY*

BY: - PANKAJ KUMAR

Group member

Akash Rauthan & Cynthia pauline

# BUSINESS PROBLEM

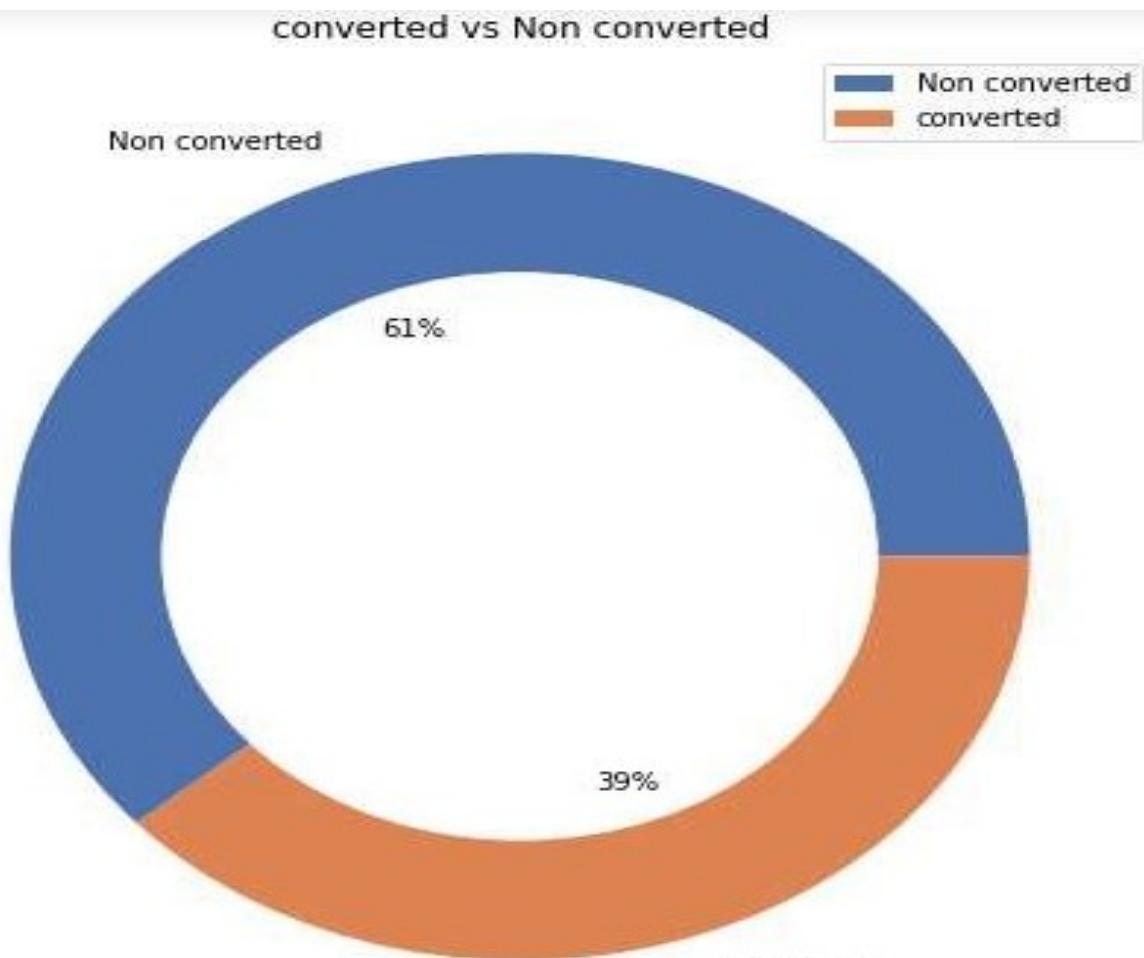
- ❖ X Education sells online courses to industry professionals.
- ❖ X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- ❖ To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- ❖ The company goals to achieve the converted rate to 80 percent . The company want to target those leads which are having a most probable chance to convert it into hot lead (i.e. those who will get converted for the course ) so that they can curtail the amount on sales and can incur more profit from it .
- ❖ It is given the target variable as "Converted" where "0" means not converted and "1" means converted .
- ❖ We will finally determine the lead score between 0 to 100 .

# SOLUTION METHODOLOGY

- ▶ To solve the given problem statement we will use the MACHINE LEARNING model in which the Logistic Regression Algorithm will helpful to determine the solution to it .
- ▶ We will import the necessary libraries and then load the data set and move a head with data sanity checking and data cleaning. After it we will do the EDA which help to know the feature and understand it well with graphs and also with relationship between them.
- ▶ We will start mapping the features and create the dummies for the categorical variable which help to convert it into the binary digit and then we will split the data into train and test data set and then we will scale the features .
- ▶ We use the RFE( Recursive feature Elimination) which is going to help for selecting the best features and than we will do the manual feature selection with the help of p value and the VIF.
- ▶ After getting all the models we will go with confusion matrix which will help to calculate specificity,sensitivity, precison and recall along with accuracy .
- ▶ Finally we will get the ROC and than we will validate the test data and we will determine the lead as per the solution.

# UNIVARIATE ANALYSIS

THE UNIVARIATE ANALYSIS EXPLAINS THE GRAPH OF THE SINGLE VARIABLE

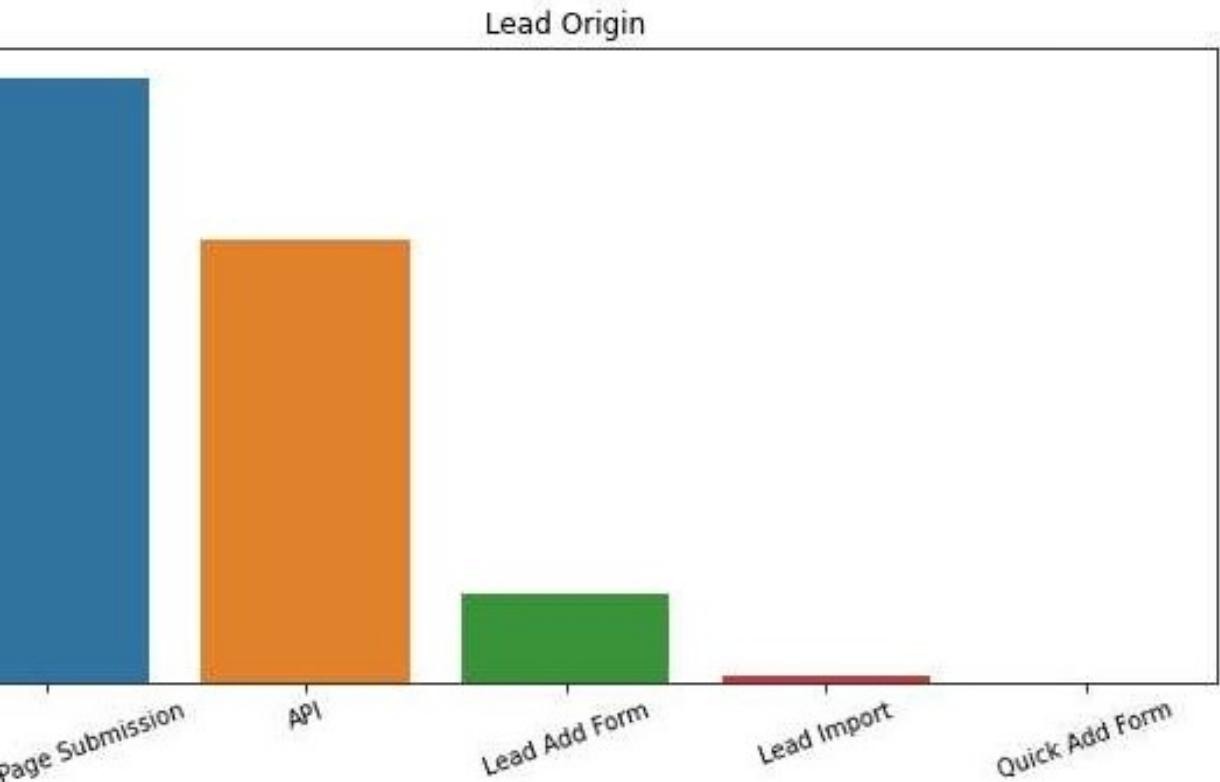


## TARGET VARIABLE : CONVERTED

As mentioned that “Converted” is the target variable which means that we are going to predict the conversion using the rest of the data set .

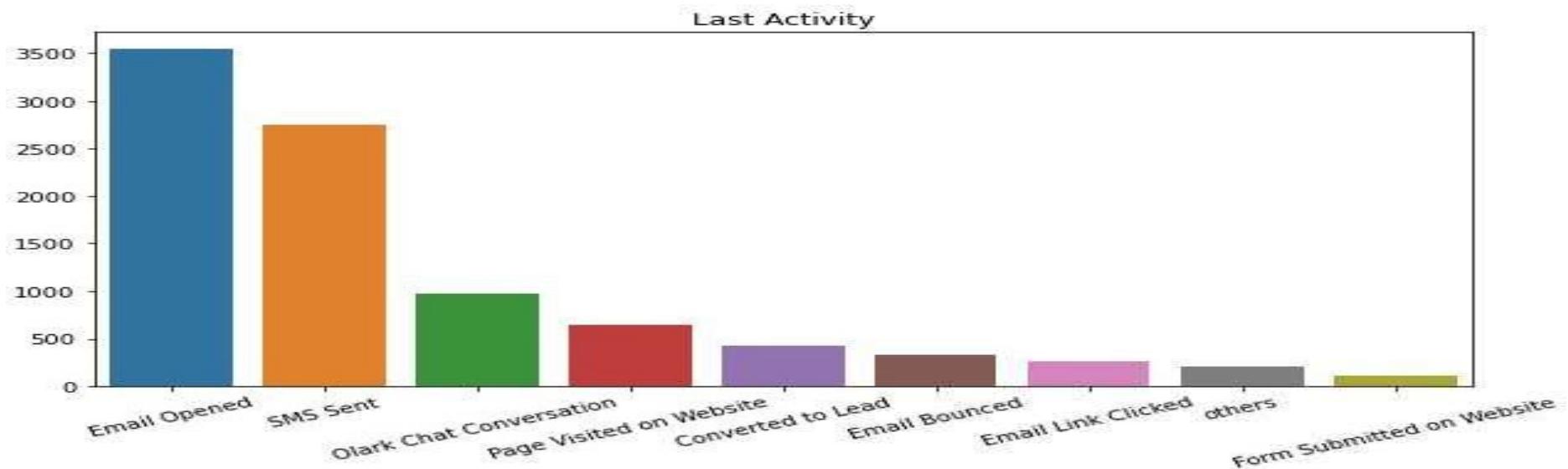
As initially we can see that the conversion rate is approximately 39%.

# LEAD ORIGIN



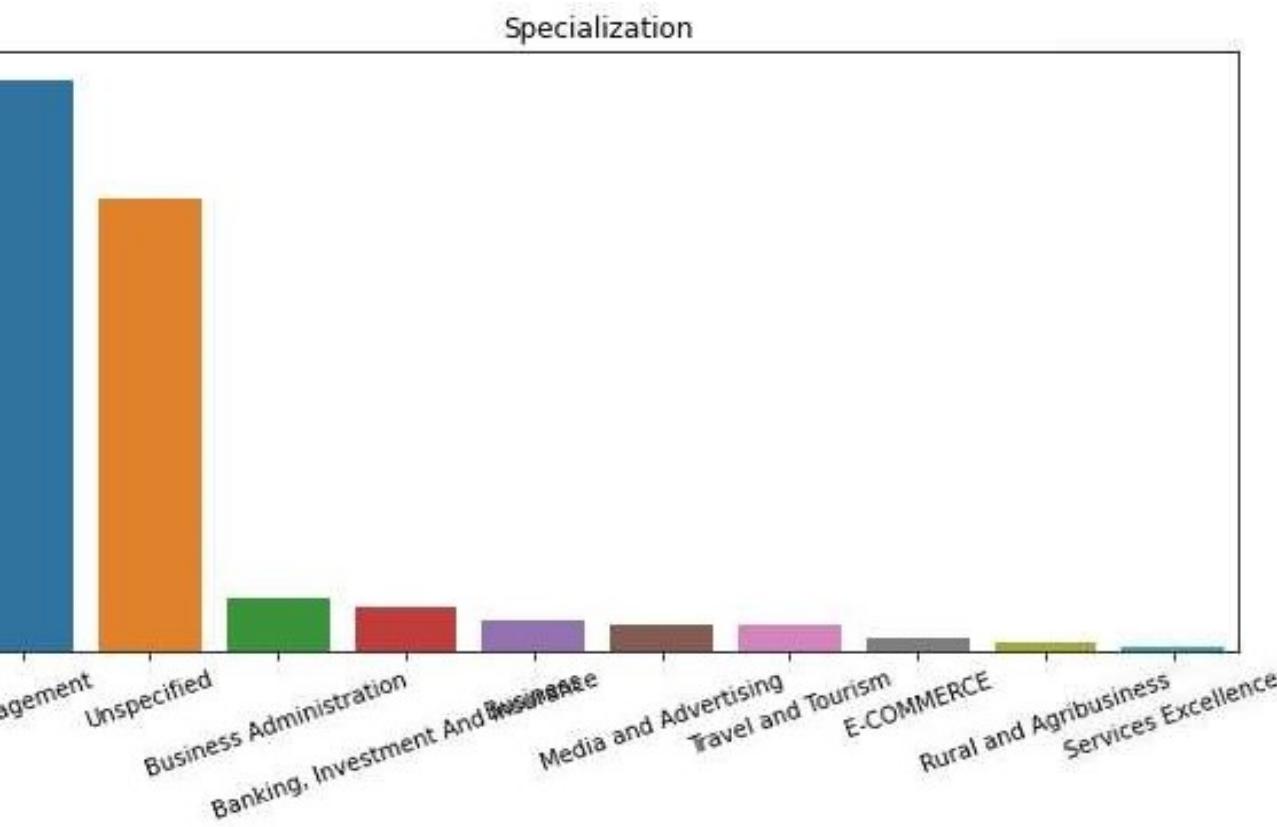
- ▶ A lead Origin act as an identifier by which the professionals are identified in different categories.
- ▶ Here, the lead origin level for the Landing page submission is higher as compare to others.

# LAST ACTIVITY



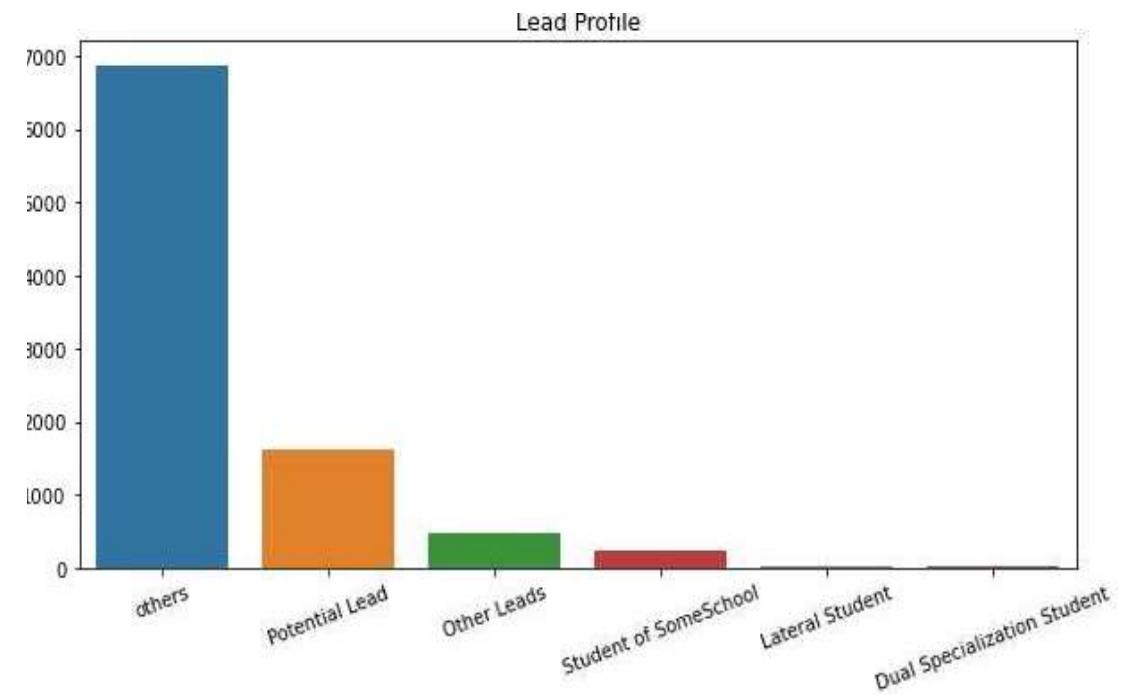
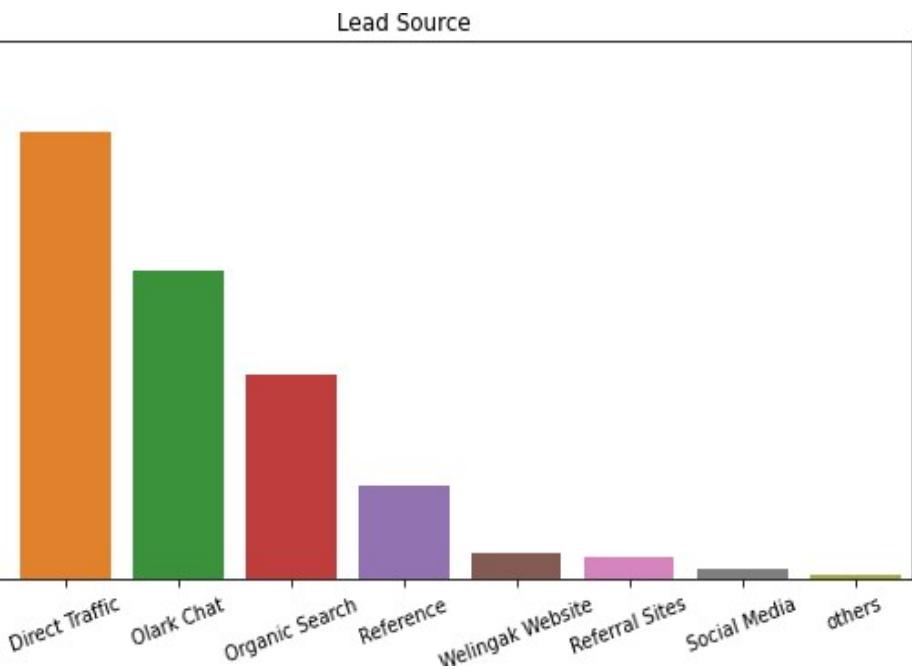
- ▶ Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.
- ▶ It means that whatever the activity is noticed at last from the lead is noted.
- ▶ The email opened played a significant role for the leads.

# SPECIALIZATION



- ▶ Specialization is a feature which tells us the background of the feature.
- ▶ Management is quite high as compare to unspecified and far more than others.

# LEAD SOURCE & LEAD PROFILE

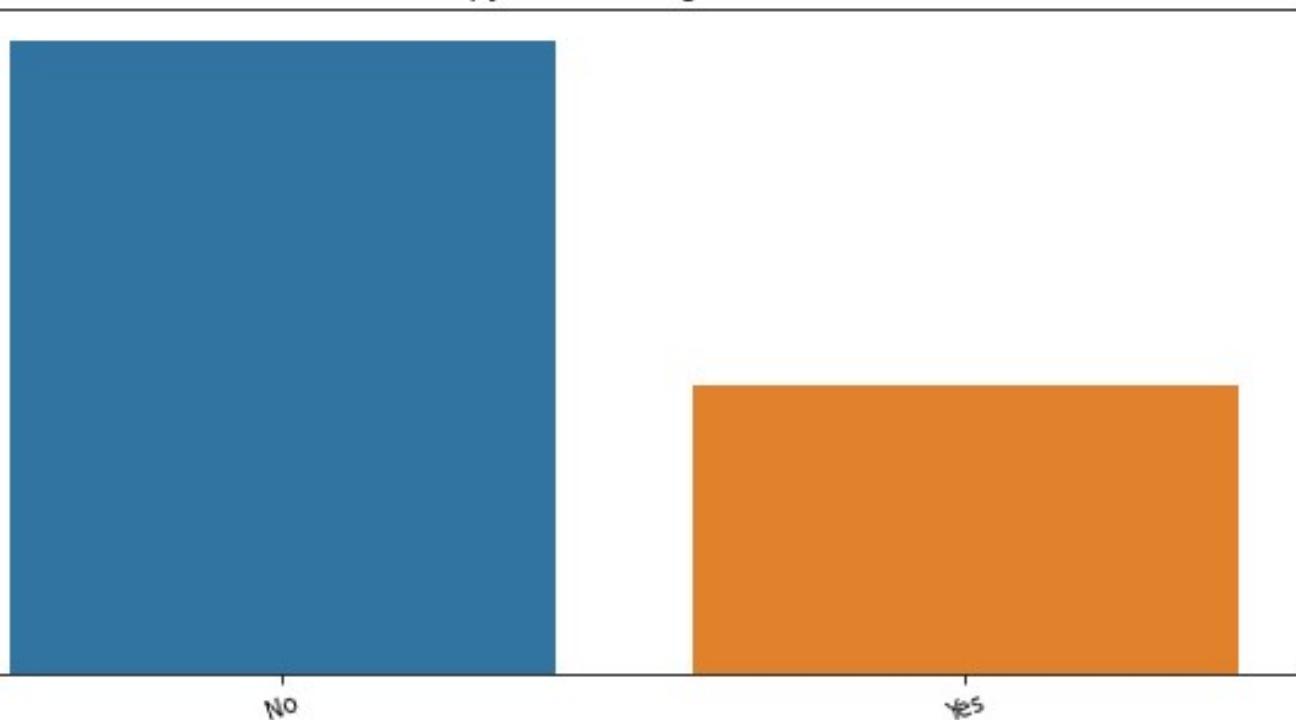


Source and Lead profile are the two different features . Lead source explains about the form from which they are coming where as Lead Profile explains the level which is provided them based on their profile.

Single is defined as the best source as others is the best category for the profile.

# FREE COPYING OF MASTERING THE INTERVIEW

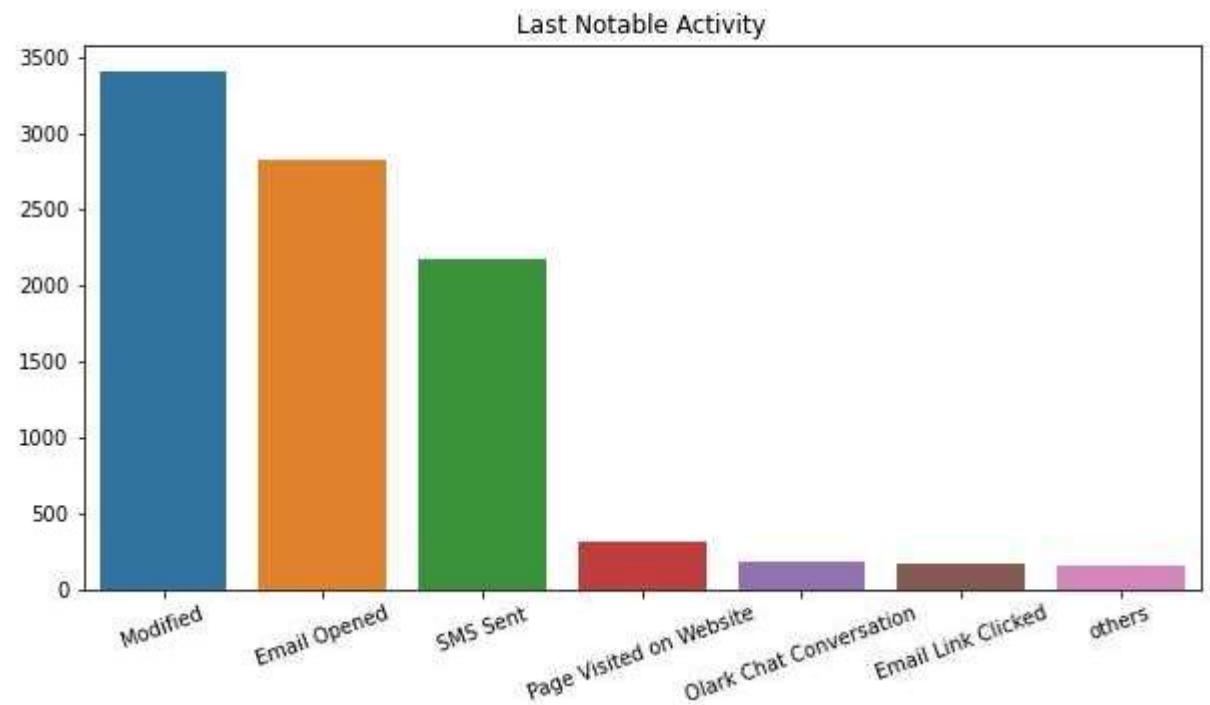
A free copy of Mastering The Interview



- ▶ A free copy of mastering the interview is given to those Leads who are being contacted by any medium.
- ▶ Most of them are answering no for it but also few are required for it .

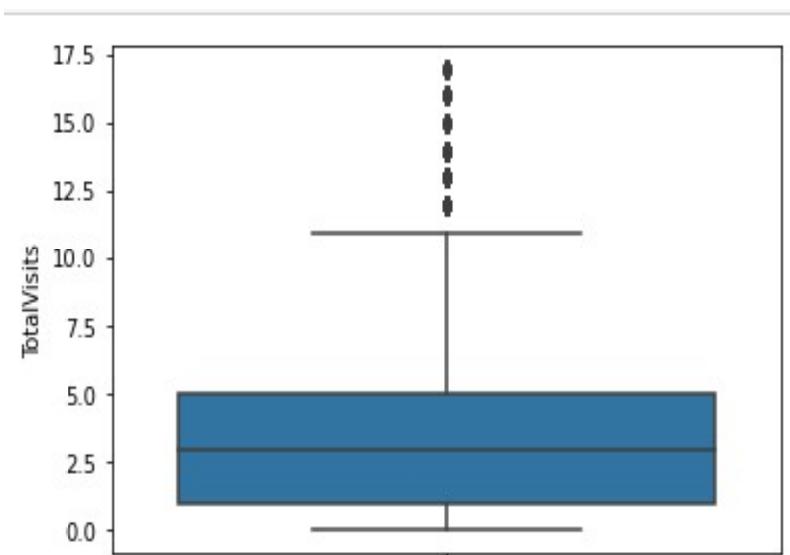
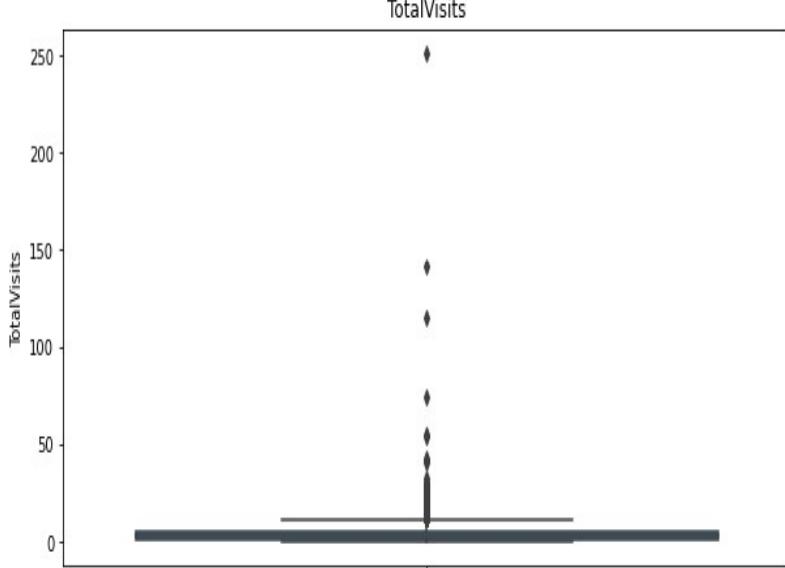
# LAST NOTABLE ACTIVITY

- ▶ What ever activity is observed by the professionals at last that is noted.
- ▶ The modified category have the highest in all follower by email opened and sms sent.



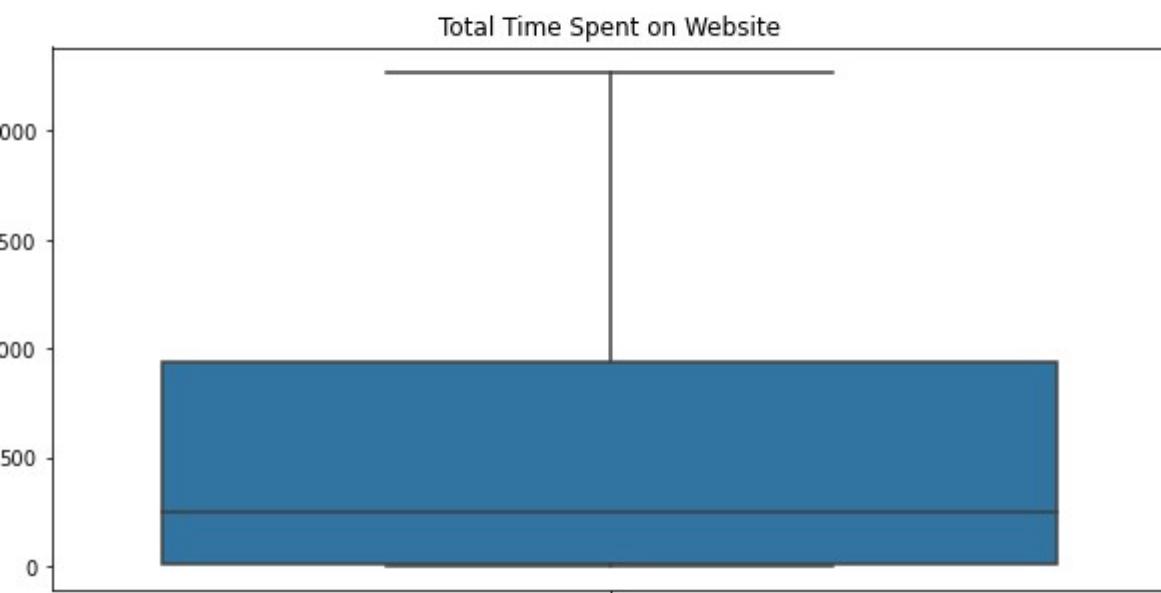
# TOTAL VISITS

Here it is seen as more outliers.



- ❖ After capping finally the box plot is clearly readable and able to see the values.
- ❖ A boxplot is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum").
- ❖ Box plot is sketched for the continuous variables only.

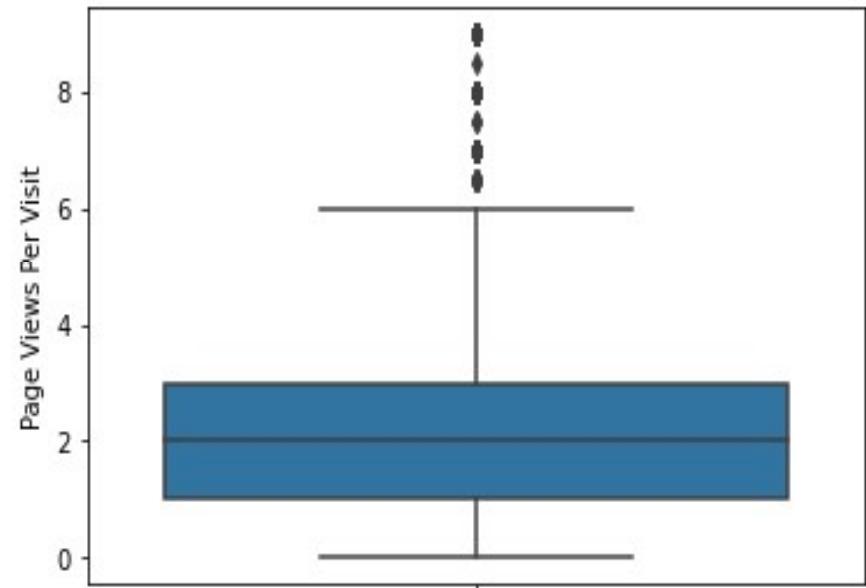
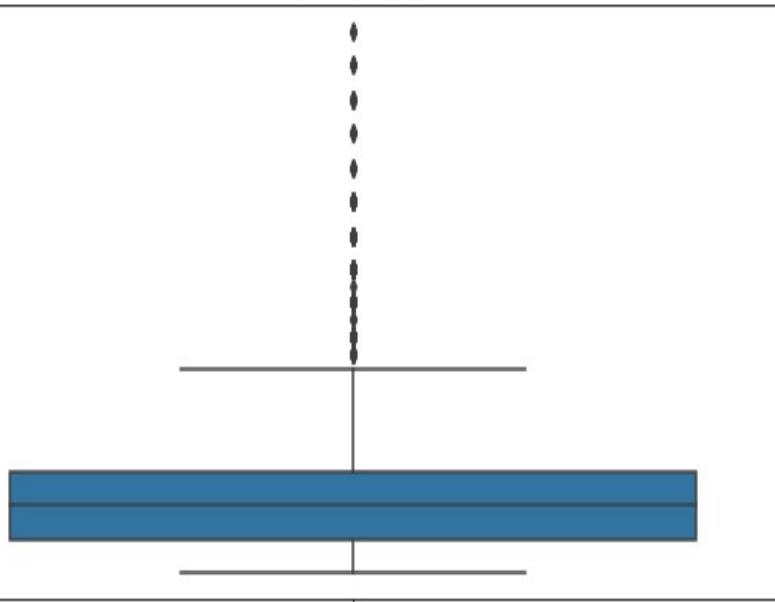
# TOTAL TIME SPENT ON WEBSITE



- ▶ As here we can observe that the plot is defined very properly with no outliers.
- ▶ As it is also to be noted that the after 3<sup>rd</sup> quartile the value reaches high .

# PAGES PER VISIT

Page Views Per Visit

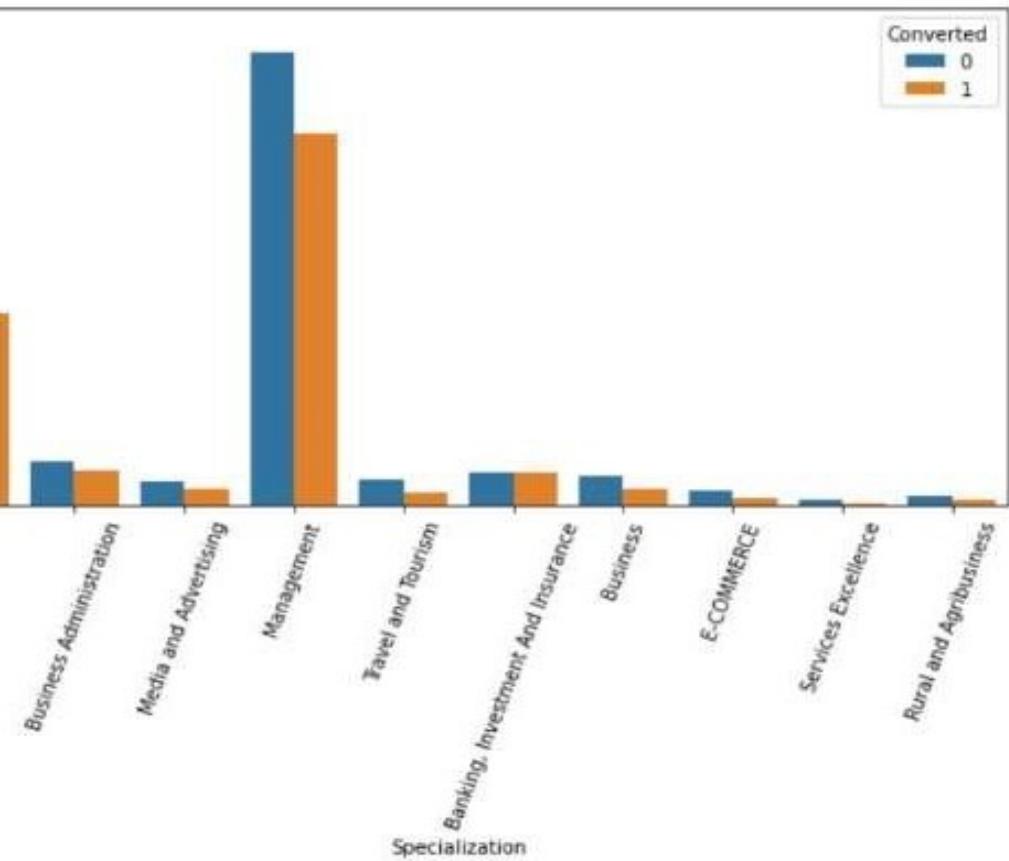


- ▶ As there are more number of outliers are seen so we do the capping at 99 percentile .
- ▶ Finally the sketch of boxplot is well defined .

# BIVARIATE ANALYSIS

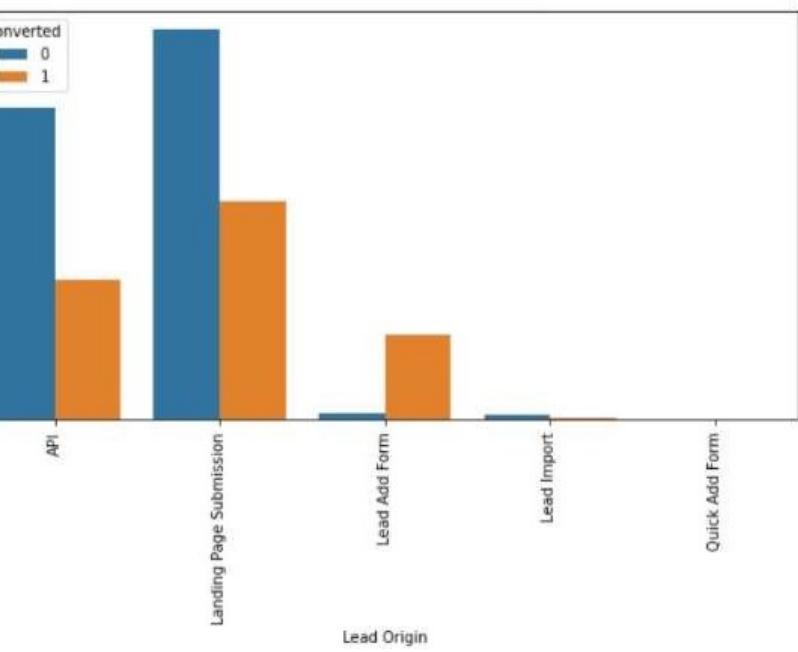
THE BIVARATE ANALYSIS EXPLAIN THE GRAPH OF THE TWO VARIABLES

## SPECIALIZATION

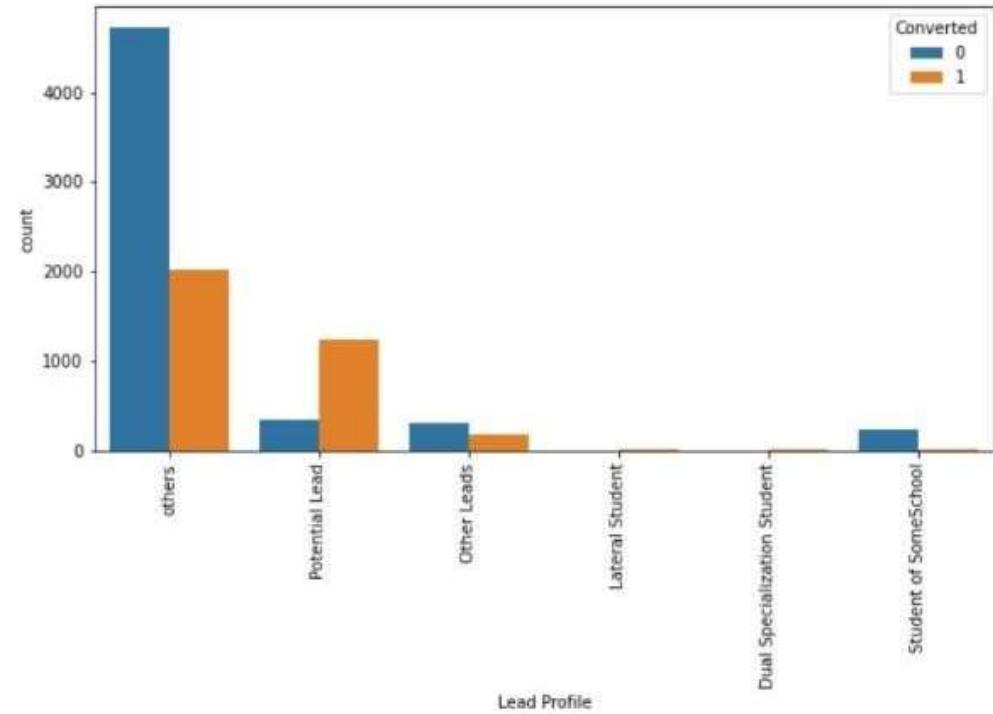


- ▶ As on the analysis we depict that the Management professional have the higher chance of converging .

# LEAD ORIGIN & LEAD PROFILE

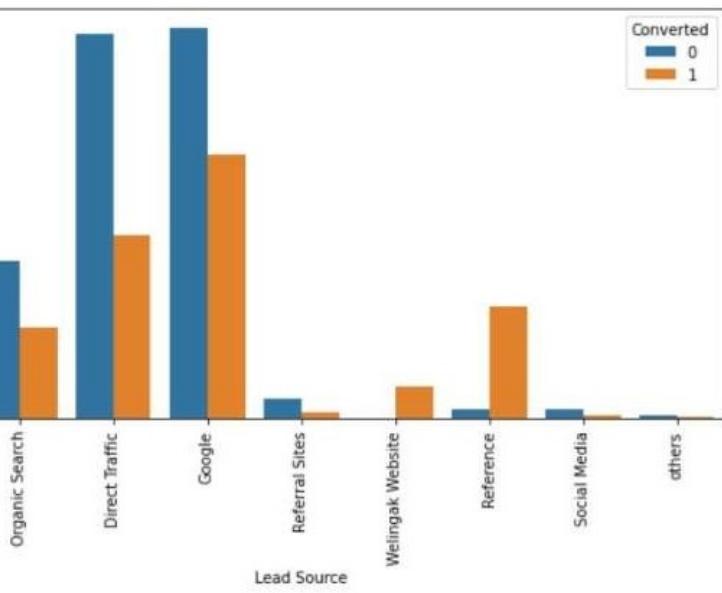


The Landing page Submission are more converting in numbers. The point to be noted that the Lead Add form have the high conversion rate as others .

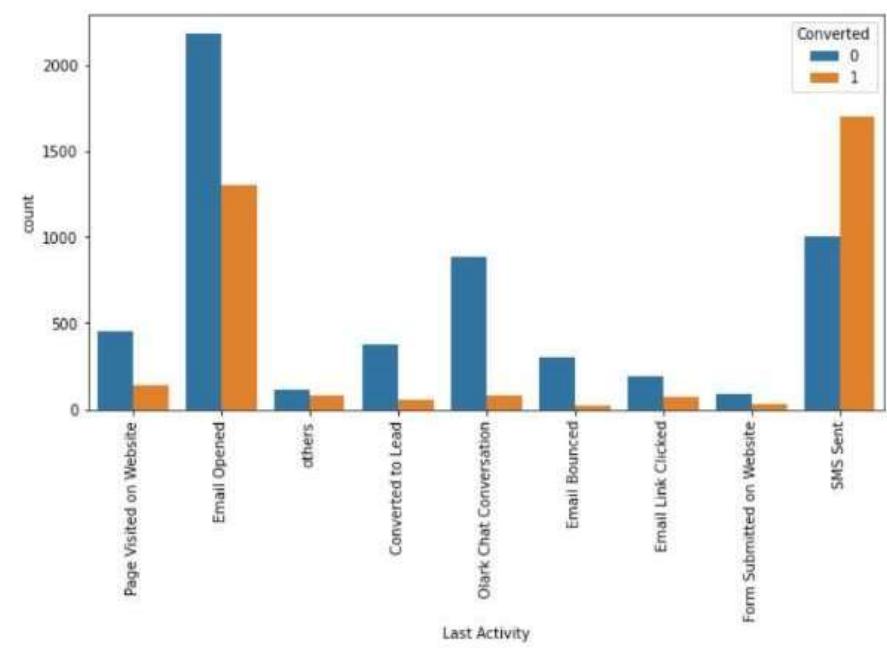


The potential lead are less in number but have high conversion rate although unspecified are more converting .

# LEAD SOURCE & LEAD ACTIVITY

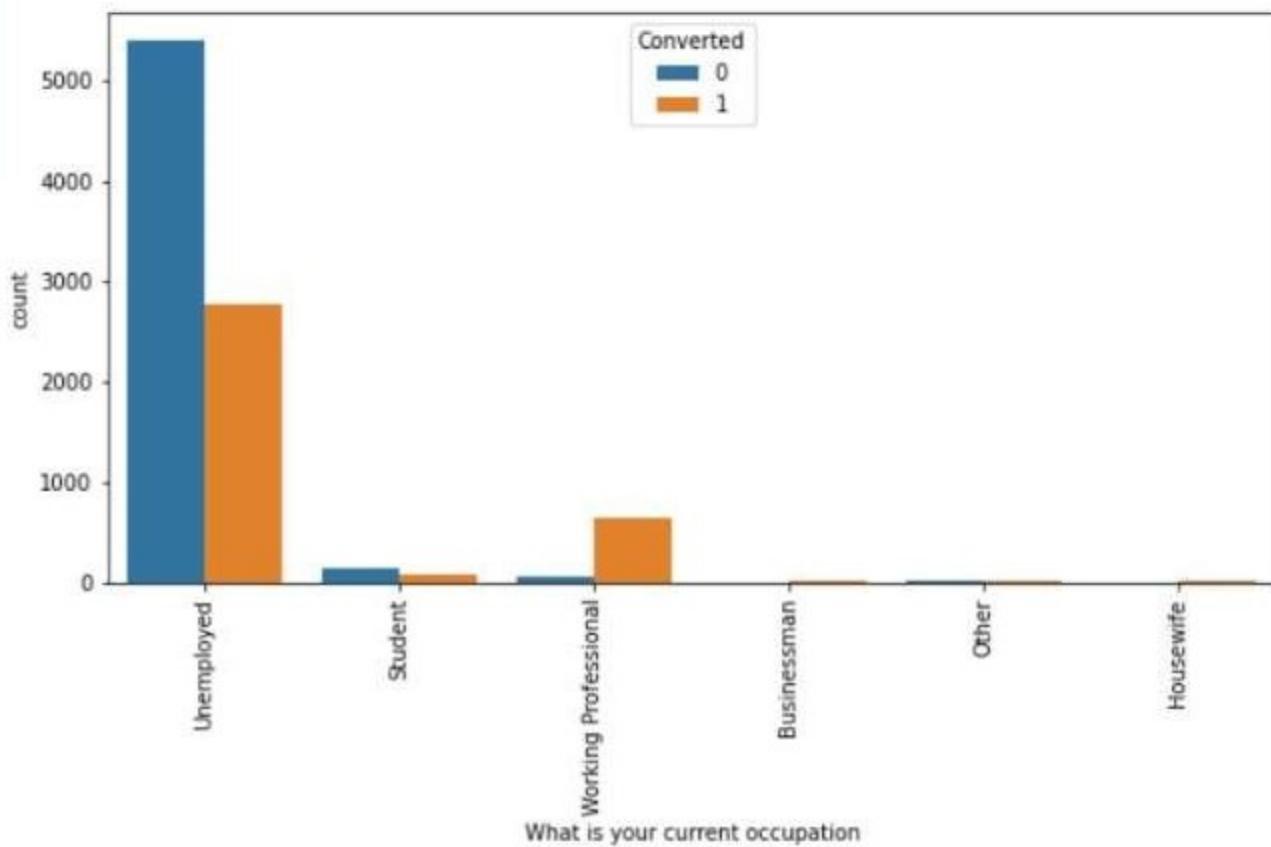


Direct traffic from Google have the most no. of conversions. Reference also have high conversion rate.



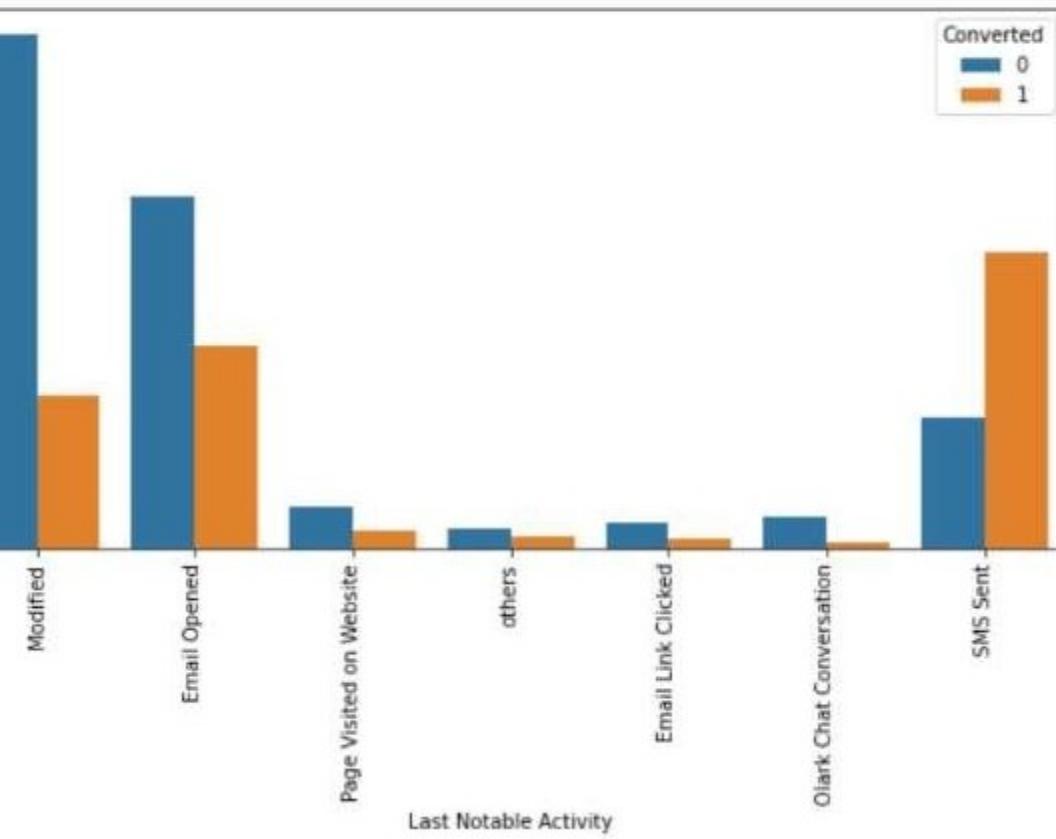
As we can see that in all bars the non conversion is high as compared to conversion. The case of SMS sent it works vice versa.

# WHAT IS YOUR CURRENT OCCUPATION



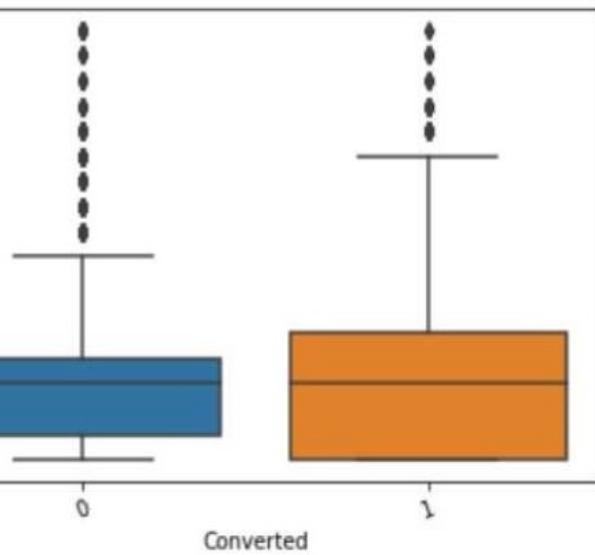
- ▶ The Unemployed category under the Current occupation has been more in numbers of getting converted as compare to others.

# LAST NOTABLE ACTIVITY



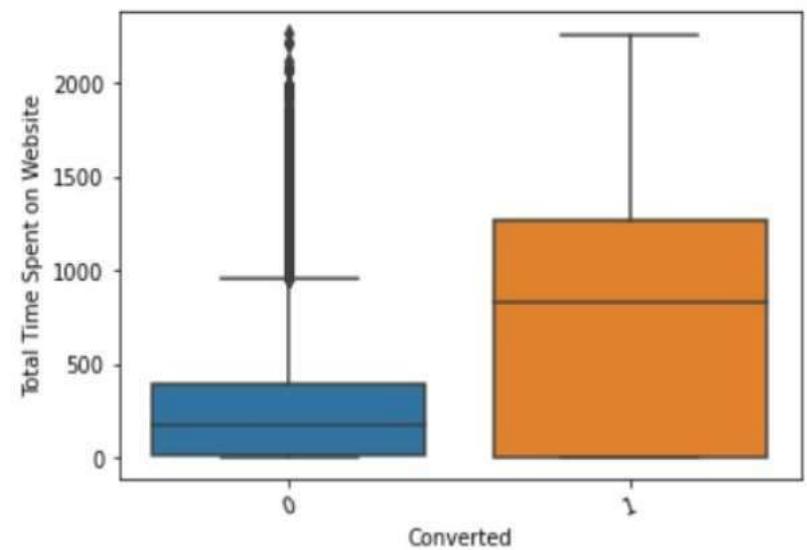
- ▶ Here it can be seen that the sms sent has the high rate of converting .
- ▶ It is also noted that it is following the same trend as Last Activity feature.

# TOTALVISITS & TOTAL TIME SPENT ON WEBSITES

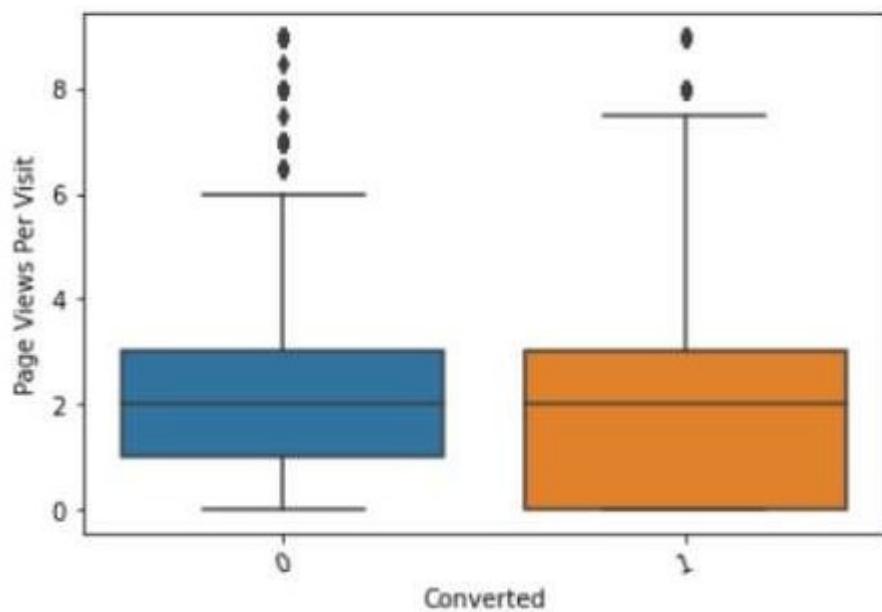


The similar trend is seen in converted and non converted but the upper and lower range differs although they have the close median value in the TotalVisits.

leads spending more time on the website are more likely to be converted. Website should be made more engaging to make leads spend more time.



# PAGES VIEWS PER VISITS



- ▶ The median value for both the converted and non converted are very close to each other so commenting on it is high risky factor.
- ▶ The range of conversion is seen high as compare to the non conversion.

# MULTIVARIATE ANALYSIS

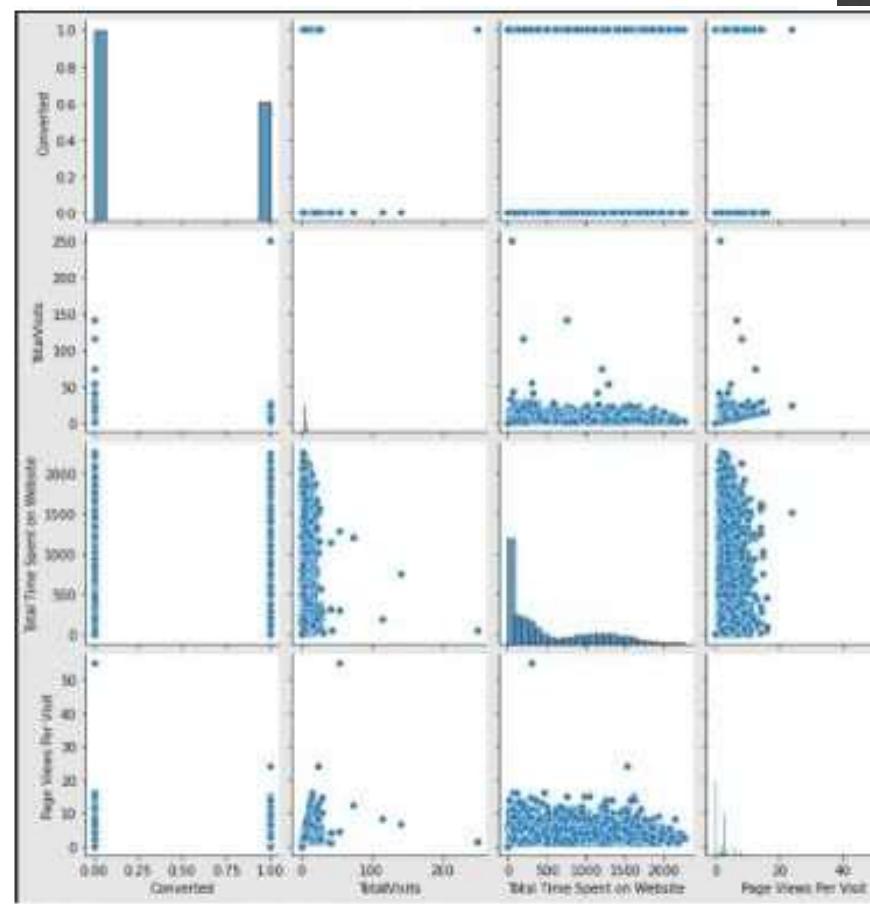
THE MULTIVARIATE ANALYSIS EXPLAINS THE GRAPH OF THE MORE THAN TWO VARIABLES.

## SCATTER PLOT

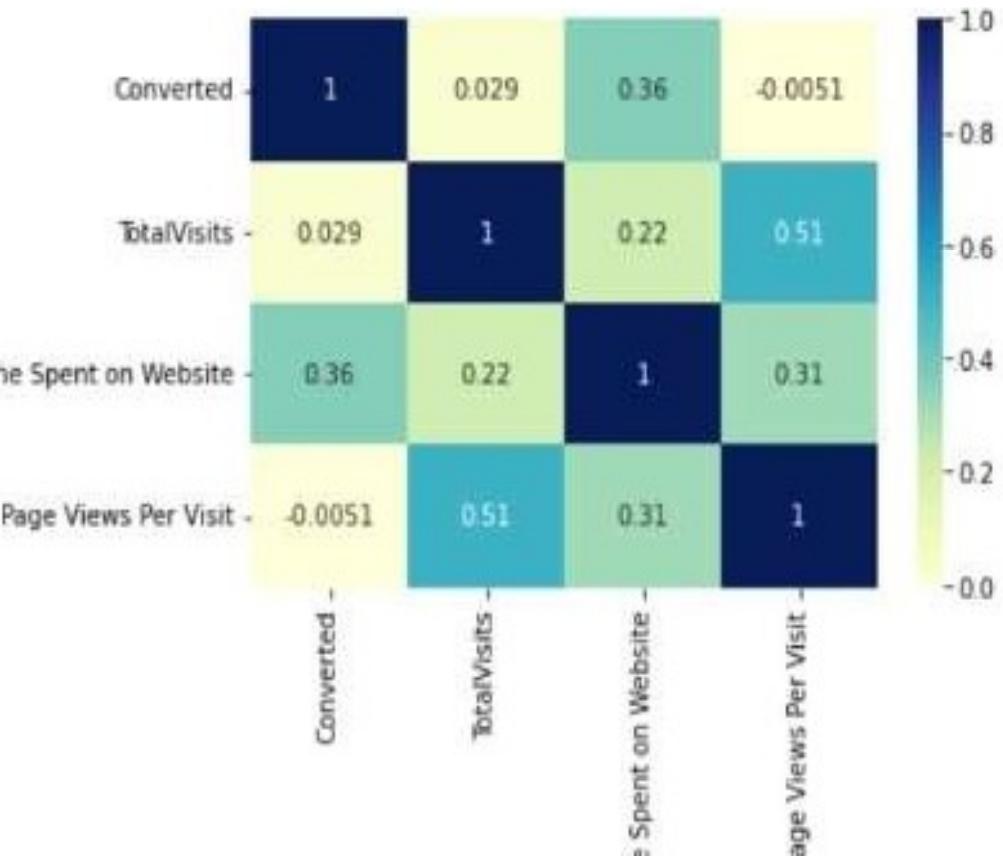
The scatter plot shows the relationship between the continuous variables.

As here we can depict that the relationship between any two variable is not well defined so there is high chance that the relationship shall be less than 0.5 which can be seen with the heatmap or with the correlation matrix.

But, eventually the relationship between the total visit and the pages per visit is seen better than others.



# HEAT MAP ON CONTINUOUS VARIABLE



The heat map defines the correlation matrix with the colour variation as the colour is as deep the matrix is more closer to 1 it shows that it is perfectly positively correlated.

It is also to be noted that relationship between 1:1 is observed in the diagonal which is always equal to 1.

Here we see that pages per visit with total visit is highly correlated whereas converted to the pages per visit is very less correlated among others.

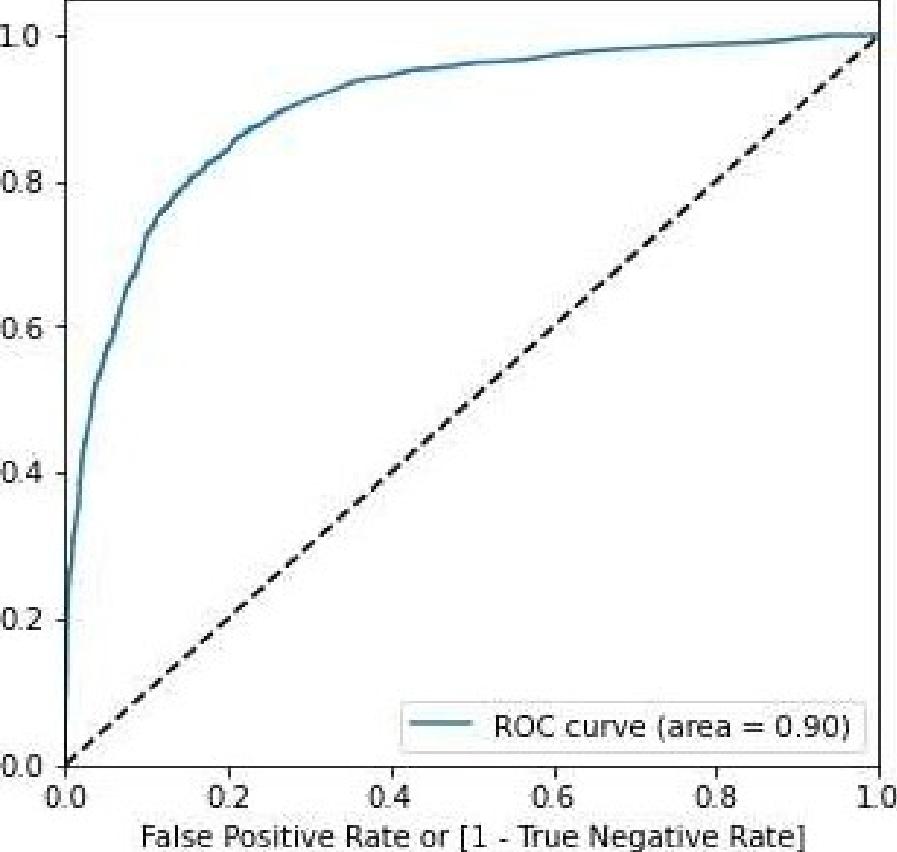
# CONFUSION MATRIX

<i>Predicted</i> -->	<i>notConverted</i>	<i>Converted</i>
<i>Actual</i>		
<i>converted</i>	3506	429
<i>notConverted</i>	631	1797

- ▶ The confusion matrix tells us how the model is correctly predicted as compared to the actual.
- ▶ It helps us to improvise the error that is to reduce the error of the prediction at some instant .

# ROC CURVE

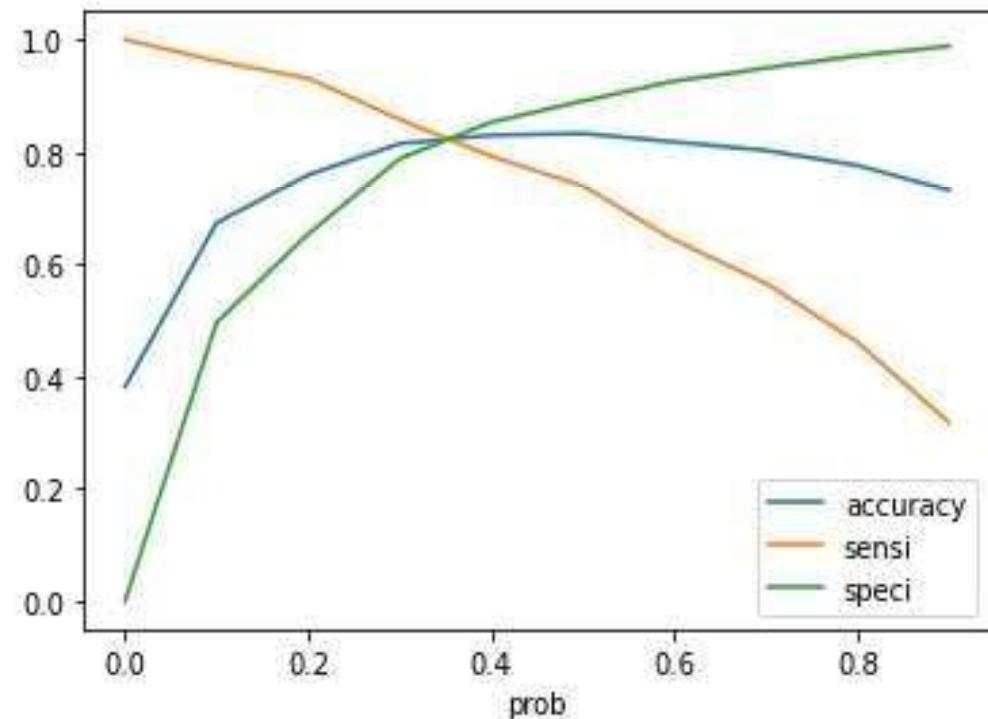
Receiver operating characteristic example



- ❖ The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or *probability of detection*. The false-positive rate is also known as *probability of false alarm* and can be calculated as (1 – specificity).
- ❖ The ROC curve is called as Receiver Operating Characteristics curve.
- ❖ As much the curve is towards the y axis the curve getting the better characteristic.

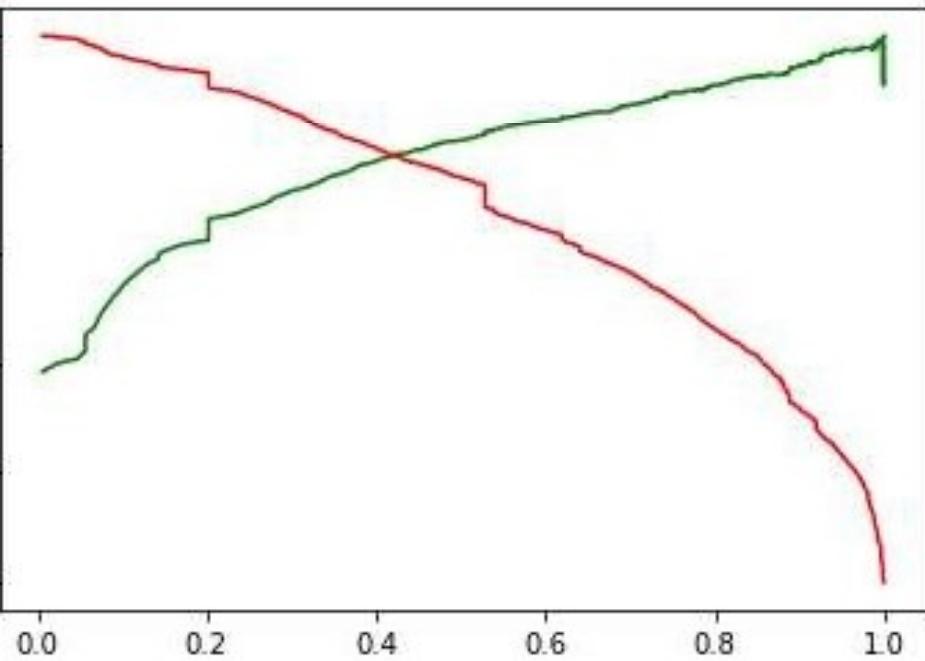
# OPTIMAL CUT OFF

- ▶ The optimal cut off is that point where the Accuracy, Specificity and the sensitivity cuts each other.
- ▶ The point of Optimal cut off is the point by which we decide the cut off prob and proceed further to calculate the value of Recall and the precision.



# THRESHOLD CURVE

---



- ▶ The threshold curve is the intersection of the precision and the recall value.
- ▶ We calculate the threshold value to determine the cut off probability .

# CONCLUSION

- ▶ Finally we have made the models with the remaining features which will help to predict the hot leads.
- ▶ This will help the CEO to reduce the cost at advertisement as well as also to gain more profit .
- ▶ The final lead score calculated lies between 0 to 100 which satisfy the problem statement.
- ▶ The model will also help us to know how much our prediction is true and how we can improvise it to get the better result with the help of the confusion matrix.

# RECOMMENDATIONS



## LEAD GENERATION

- It is good to collect data often and run the model and get updated with the potential leads. There is a belief that the best time to call your potential leads is within few hours after the lead shows interest in the courses.
- Along with phone calls, it's good to mail the leads also to keep them reminding as email is as powerful as cold calling.
- Reducing the number of call attempts to 2-4 and increasing the frequency of usage of other media like advertisements in Google, or via emails to keep in touch with the lead will save a lot of time.
- Focusing on Hot Leads will increase the chances of obtaining more value to the business as the number of people we contact are less but the conversion rate is high.

**THANK YOU**