LEAD SCORE CASE STUDY

Prepared by:

Nisha Rao

Neha Jhunjhunwala

Dileep Kumar B P

Problem statement

- Ed-tech company named X Education sells online courses to industry professionals. Many interested candidates land on their websites to get more details.
- Once, interested person fills basic details. X Education, gets a data for leads, its lead conversion rate is very poor approximately 30%.
- To make good business out of it, the company wants to identify the most potential leads, also known as ‘Hot Leads’.
- If they are able to identify this set of leads, the lead conversion rate increases as the sales and marketing team will now be focusing more on communicating with the potential leads instead of making calls to everyone.

Objective:

- The company wants to identify the most potential leads.
- Have to build a Model which identifies the hot leads.
- For future use user cases need to generate a model.

Steps followed:-

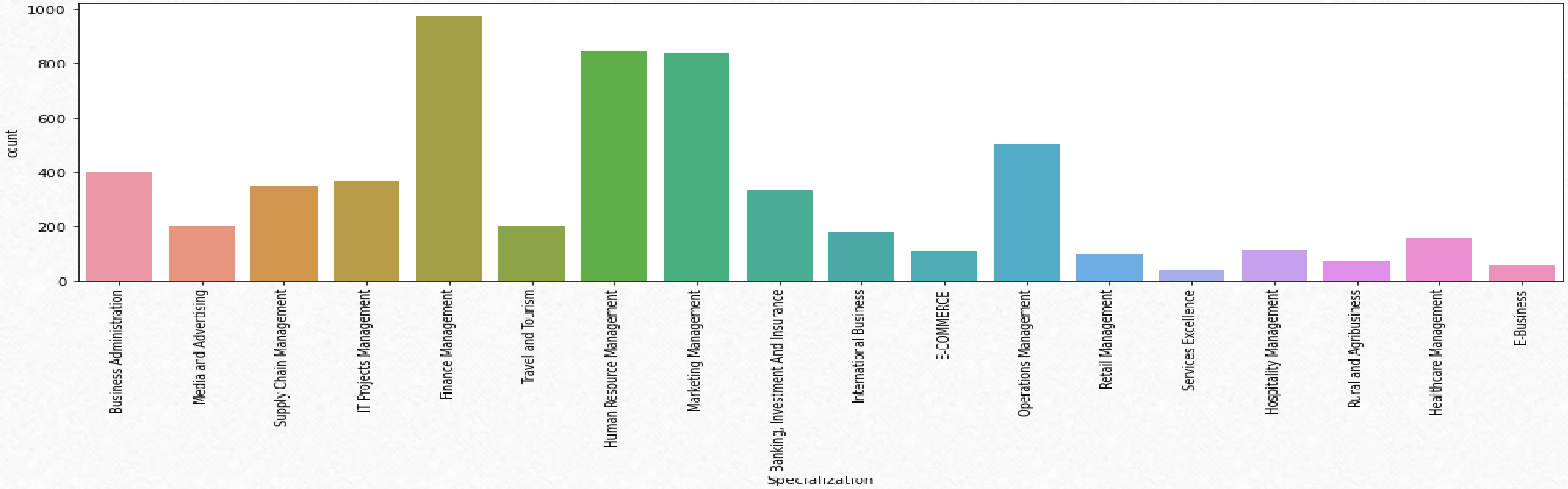
- Data Inspection.
- Data Cleaning.
- Exploratory Data Analysis.
- Univariate Analysis and Bivariate Analysis.
- Data Preparation
  1. Converting some binary variables
  2. Creating Dummy variables for the categorical features
  3. Splitting the data into train and test set
  4. Scaling the features
- Feature Selection Using RFE.
- Model Building
- Checking for P-values and VIF values.
- Making Prediction on the Train set
- Choosing an arbitrary cut-off probability point of 0.5 to find the predicted labels.
- Making the Confusion matrix.
- Plotting the ROC Curve
- Finding Optimal Cut off Point
- Model Evaluation
- Precision and Recall
- Making predictions on the test set
  1. Scaling the test data
  2. Assigning Lead Score to the Testing data
- Recommendations

Main raw dataset:-

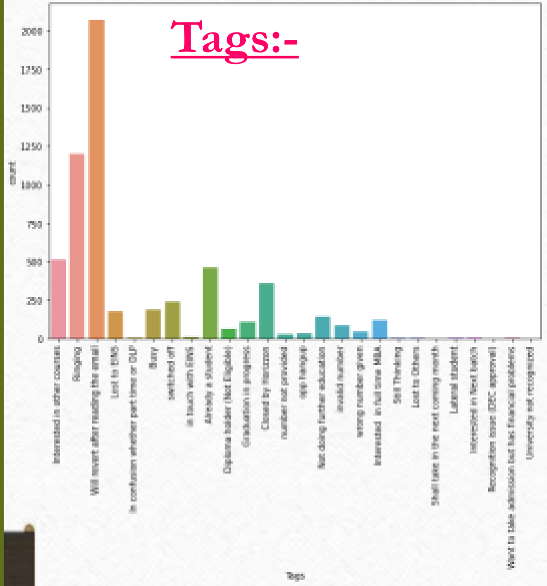
- Initial data is having 9240 rows and 37 columns.
- All datatypes are in correct format checked.
- We dropped the columns with missing values greater than 40%.(i.e. 'How did you hear about X Education','Lead Quality',’Lead Profile','Asymmetrique Activity Index','Asymmetrique Profile Index','Asymmetrique Activity Score', 'Asymmetrique Profile Score’).

Specialization

- There is 37% missing values present in the Specialization column .
- 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.
- So we can create a another category 'Others' for this.

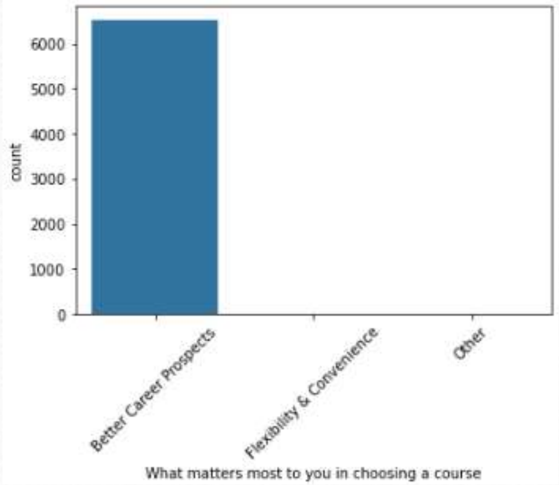


Tags:-



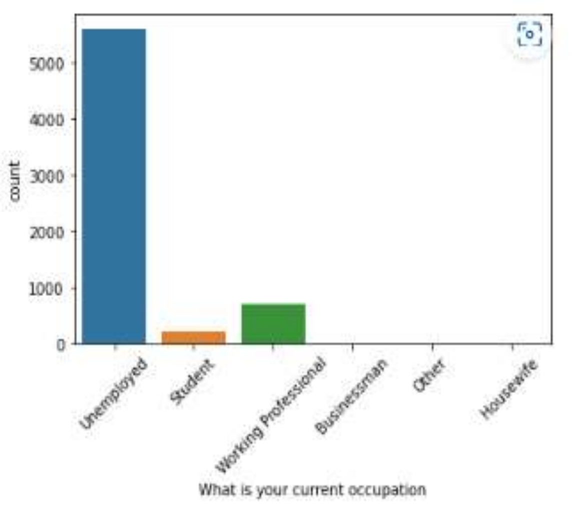
- ➢ More values are 'Will revert after reading the email' ,. So we need to target this variable

What matters most to you in choosing a course:-



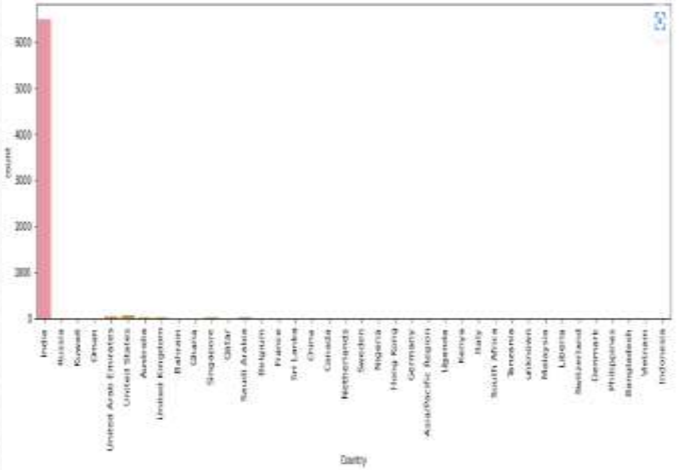
- ➢ Most of the values are in the better career prospectus.
- ➢ So, we can see that this is highly skewed column.

'What is your current occupation :-



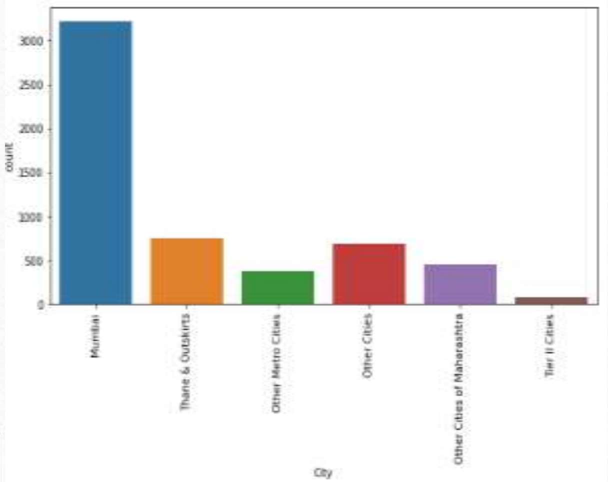
- ➢ Most values are in column. 'Unemployed'. Followed by working professional and student.

Country:-



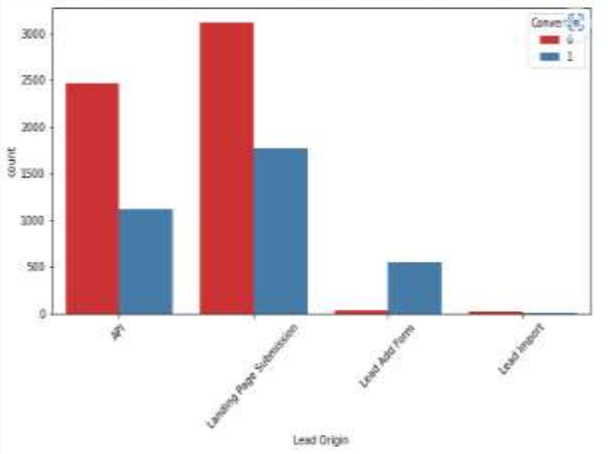
- ➢ We can see that this is highly skewed column but it is an important information with respect to the lead. Since most values are 'India' we need to target on it.

City:-



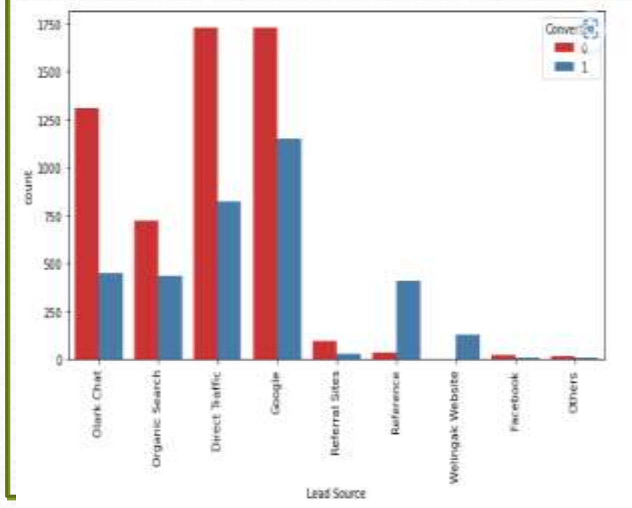
- ➢ We can see large number of leads from Mumbai followed by other cities.

Lead Origin:-



- ➢ API and Landing Page Submission have mediocre conversion rate but count of lead originated from them are considerable.
- ➢ Lead Add Form has very high conversion rate but count of leads are not very high.
- ➢ Lead Import are very less in count.

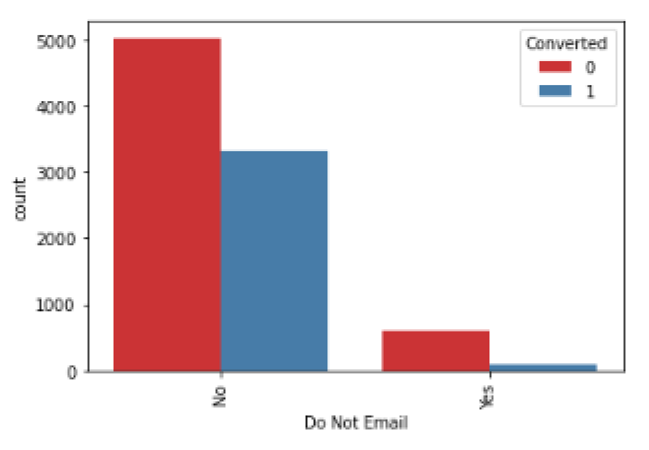
Lead Source:-



- ➢ Google and Direct traffic generate maximum number of leads.
- ➢ Conversion Rate of reference leads and leads through Welingak website is high.

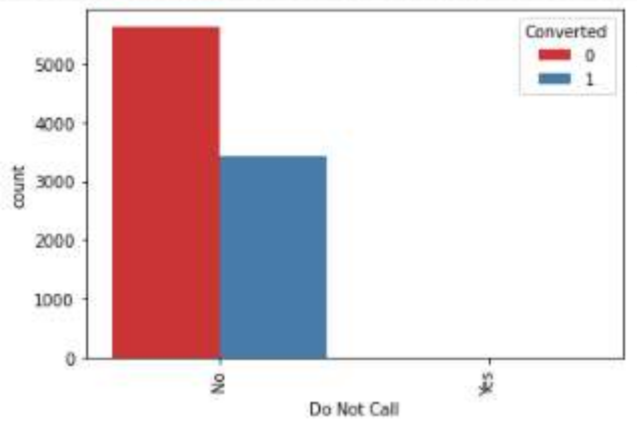
.

Do not Email:-



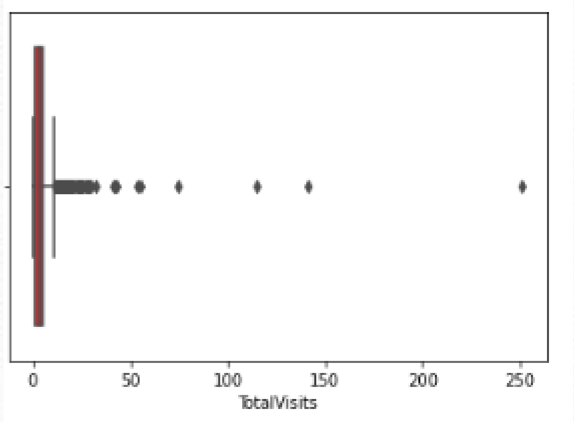
- ➢ Most entries are 'No'. No Inference can be drawn with this parameter.

Do not call:-



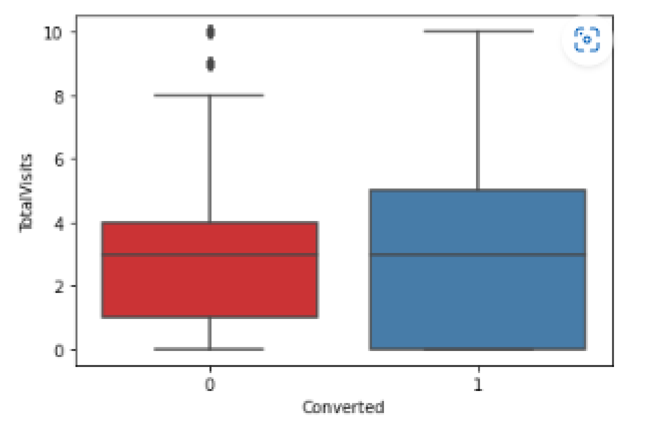
- ➢ Most entries are 'No'. No Inference can be drawn with this parameter

Total Visits:-



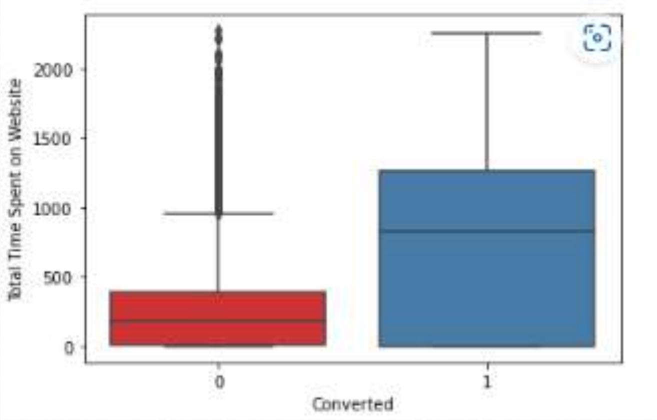
- ➢ we can see there are a number of outliers in the data.

Total Visits :-



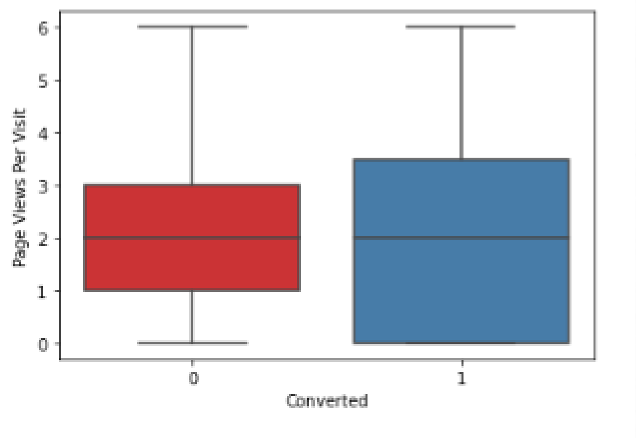
- ➢ Google and Direct traffic generates maximum number of leads.
- ➢ Conversion Rate of reference leads and leads through welingak website is high.

Total Time Spent on Website:-



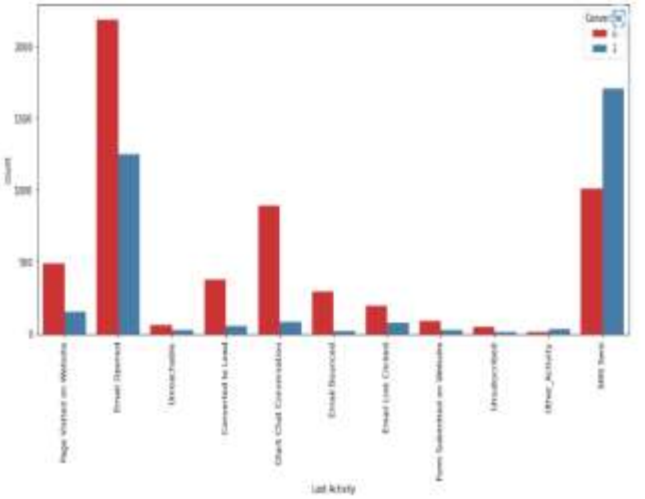
- ➢ Leads spending more time on the weblise are more likely to be converted.
- ➢ Website should be made more engaging to make leads spend more time**.**

Page Views Per Visit:-



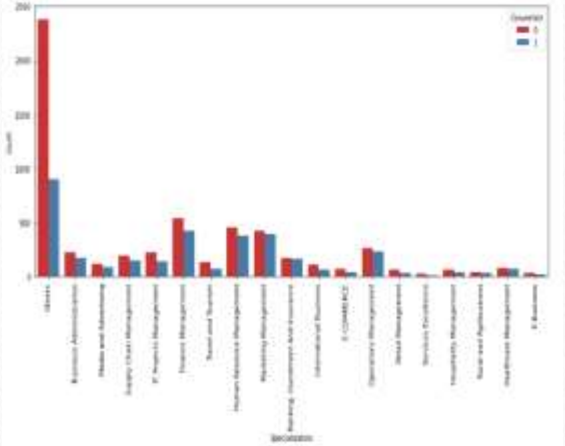
- ➢ Median for converted and unconverted leads is the same.
- ➢ Nothing can be said specifically for lead conversion from Page Views Per Visit.

Last Activity:-



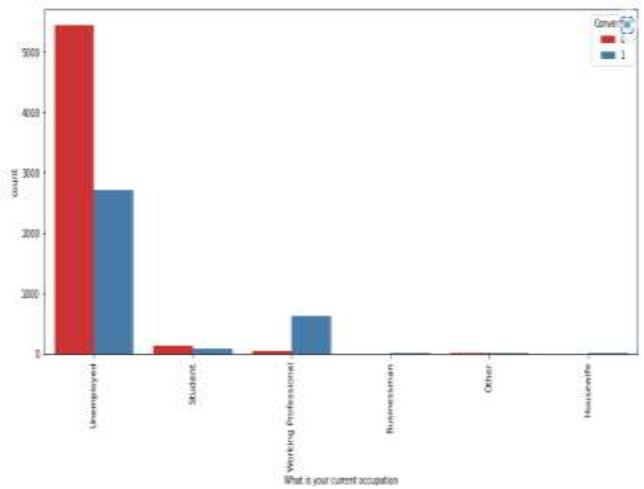
- ➢ Most of the leads have their Email opened as their last activity.
- ➢ Conversion rate for leads with last activity as SMS Sent is relatively pretty high.

Specialization:-



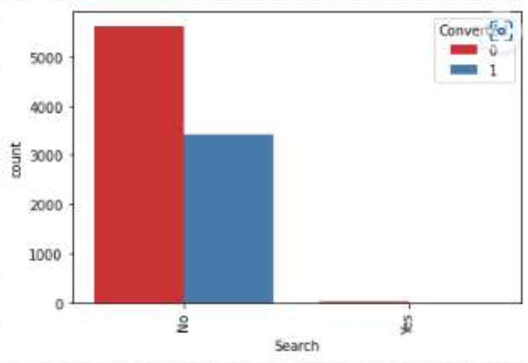
- ➢ Large ration of conversion rate we can see for different streams of management.

What is your current occupation:-

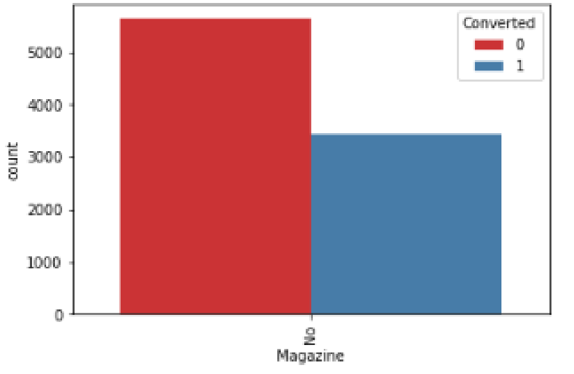


- ➢ Working Professionals going for the course have high chances of joining it.
- ➢ Unemployed leads are the most in numbers but have relatively lower conversion rate.

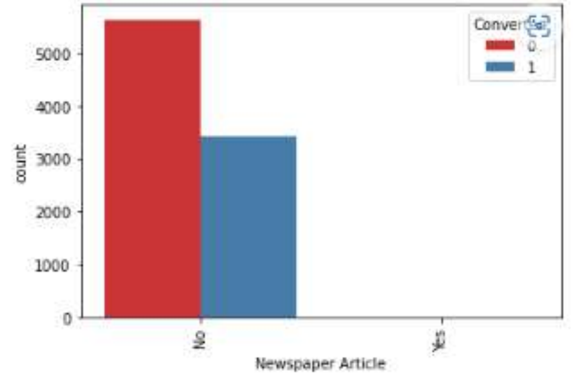
Search:-



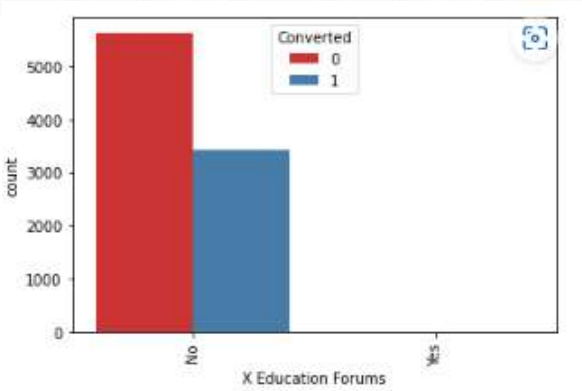
Magazine:-



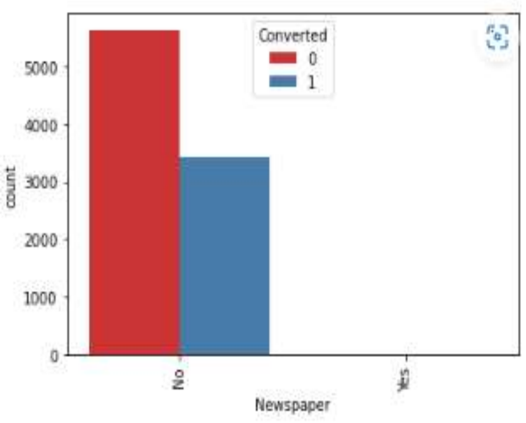
Newspaper Article:-



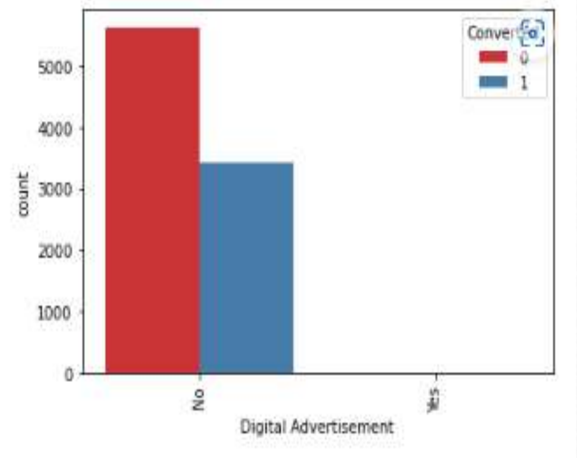
X Education Forums:-



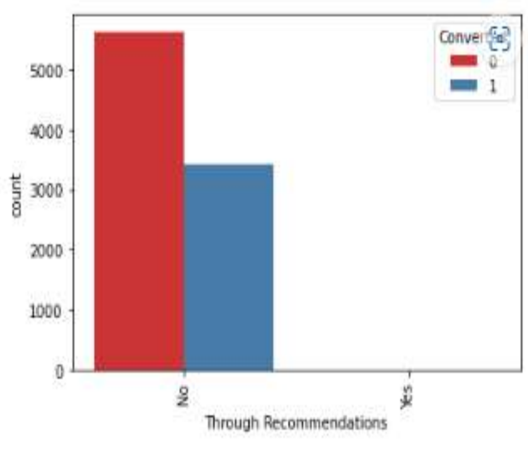
Newspaper:-



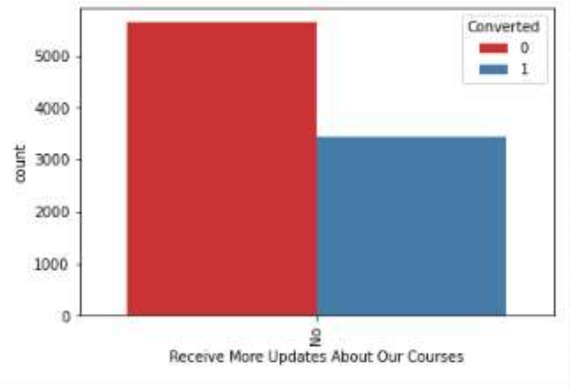
Digital Advertisement:-



Through Recommendations:-

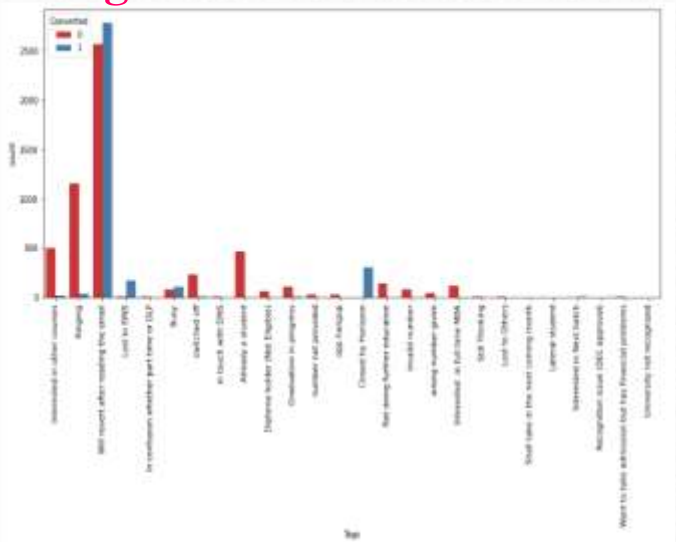


Receive More Updates About Our Courses:-



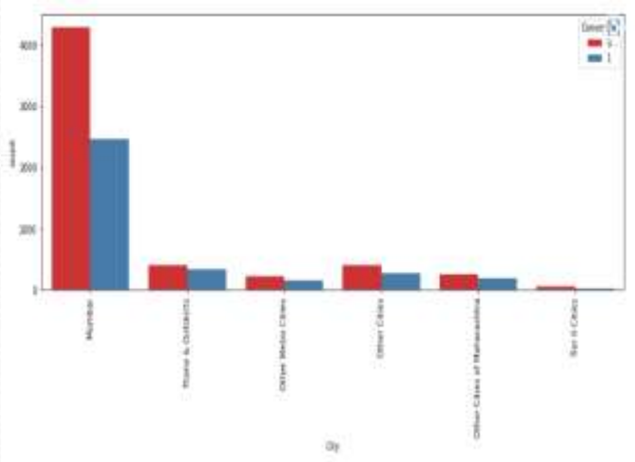
- Most entries are 'No'. No Inference can be drawn with this parameter.

Tags:-



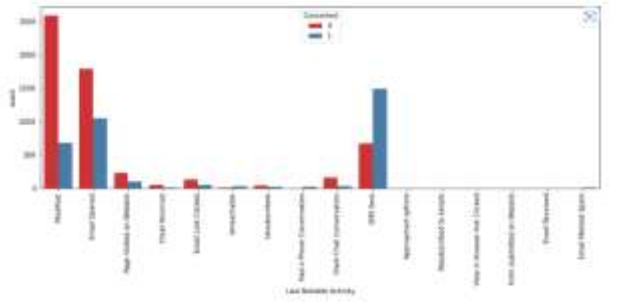
- Since this is a column which is generated by the sales team for their analysis , so this is not available for model building . So we will need to remove this column before building the model..

City:-



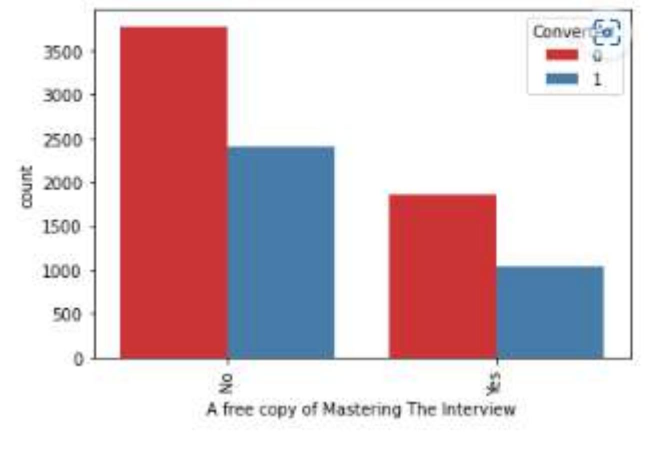
- Most leads are from Mumbai with around 50% conversion rate.

Last Notable Activity:-



- SMS sent is having more rate of conversion value compared to other.

A free copy of Mastering The Interview



- Most entries are 'No'. No Inference can be drawn with this parameter.

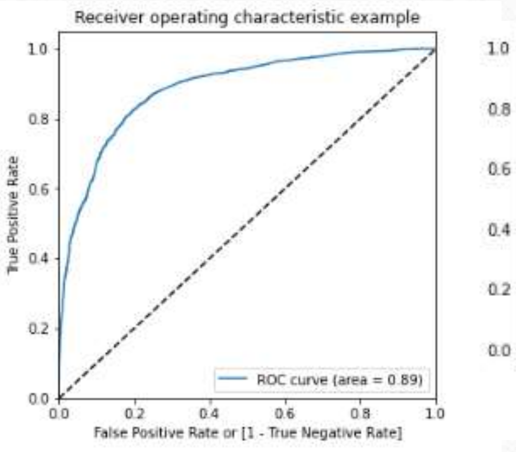
Model Building:-

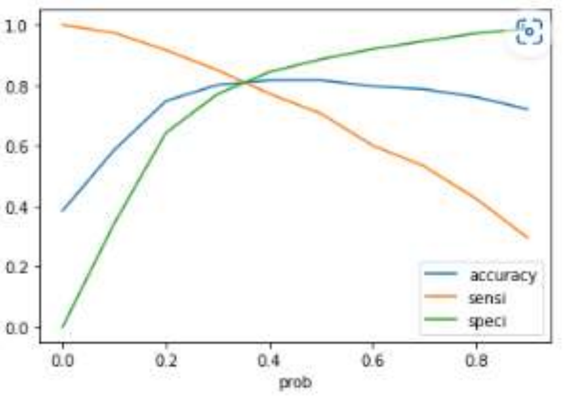
- Converting some binary variables
- Creating Dummy variables for the categorical features
- Dropping the columns for which dummies were created
- Splitting the Data into Train and Test Sets.
- Scaling the features
- Feature Selection Using RFE.
- Assessing the model with Stats Models-Drop the columns with high P-value and VIF values. Model-9 is our final model. We have 12 variables in our final model.
- Making Prediction on the Train set
- Creating a dataframe with the actual Converted flag and the predicted probabilities.
- Making the Confusion matrix
- Metrics beyond simply accuracy
- Plotting the ROC Curve
- Finding Optimal Cutoff Point
- Assigning Lead Score to the Training data.

Model Evaluation:-

- **Overall accuracy ~81%**

ROC curve:-

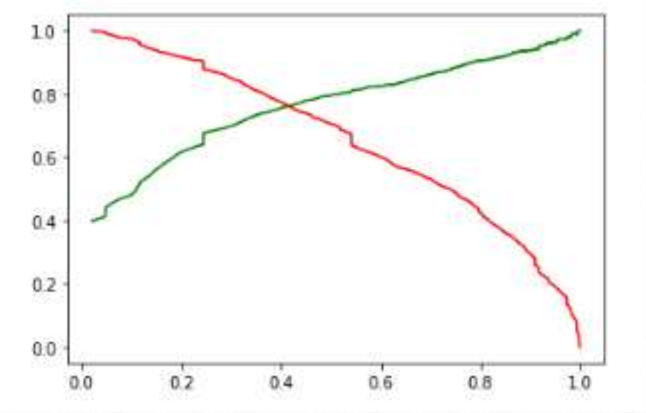




- we have higher **(0.89**) area under the ROC curve , therefore our model is a good one.
- From the curve above**, 0.34** is the optimum point to take it as a cutoff probability

Precision and recall trade-off:-

- **Precission~79%**
- **Recall ~70**.



The above graph shows the trade-off between the Precision and Recall

Observations:

After running the model on the Test Data , we obtain:

- **Accuracy : 80.4 %**
- **Sensitivity : 80.4 %**
- **Specificity : 80.5 %**

Results :

1) Comparing the values obtained for Train & Test:

Train Data:

- **Accuracy : 81.0 %**
- **Sensitivity : 81.7 %**
- **Specificity : 80.6 %**

Test Data:

- **Accuracy : 80.4 %**
- **Sensitivity : 80.4 %**
- **Specificity : 80.5 %**

Thus we have achieved our goal of getting a ballpark of the target lead conversion rate to be around **80%** . The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model to get a higher lead conversion rate of **80%**.

2) Finding out the leads which should be contacted

- ➢ The customers which should be contacted are the customers whose "Lead Score" is equal to or greater than 85. They can be termed as 'Hot Leads'.
- ➢ So there are 368 leads which can be contacted and have a high chance of getting converted. The Prospect ID of the customers to be contacted are

3) Finding out the Important Features from our final model

Recommendations:

- ➢ 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.

Thankyou