# LEADS SCORE CASE STUDY SUBMISSON REPORT

Group by :
SRI SAIRAM
MONOJ DAS
MANISHA SINGH

# BUSINESS OBJECTIVE

To build logistic regression model to predict weather a lead for online courses offered by X education would converted successfully or not.

Here we need to help X education to find out there Hot leads that can be easily converted to paying customers. To build logistic regression model for this we need to assign lead score values between 0 to 100 to each of the leads which can be used by the company as potential lead targets.

To achieve are aim there are sub goals as listed below :

- Create a logistic regression model to predict lead conversion probabilities.
- To get a threshold above which the leads predicated to be converted whereas not converted if it is below it.
- Multiply the lead conversion probability to arrive at the lead score value to each lead.

# STEPS INVOLVED

**There are the following steps involved to achieve our objective :**

- Read and understanding of the data
- Data cleaning
- Prepare data for model building
- Model building
- Model Evaluation
- Making predictions Training split tests and Test split tests

# Reading and understanding data

- First all import required libraries and suppress all the warnings.
- Import the data .csv file.
- *Fetch the first few entries.*
- *Inspect the shape, size , description of the dataset.*
- *Inspect the different columns of the dataset.*
- *Check the information of the data like data types of each columns, statistical feature numerical columns and presence of null values in different columns.*
- Check the duplicity of the data .

# DATA CLEANING

## Checking and imputing Missing values and Null values

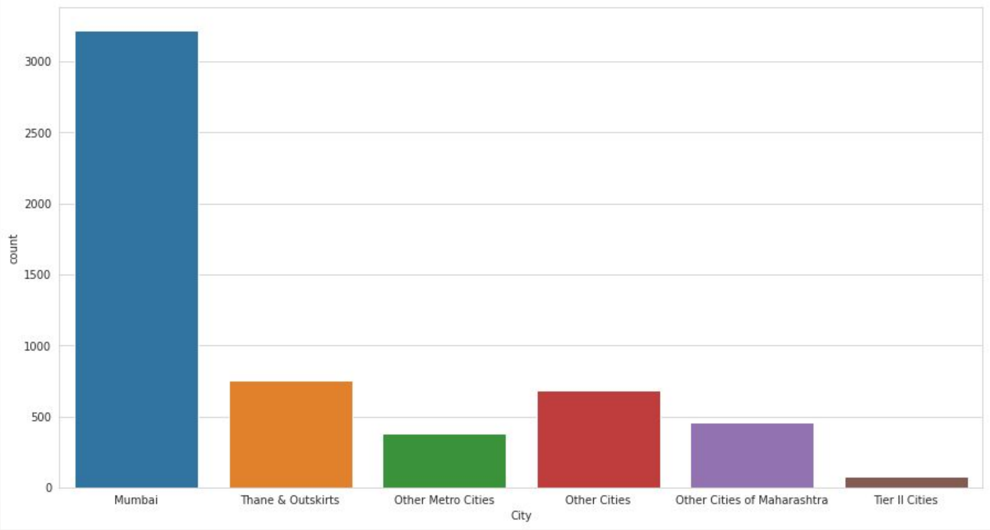
- Replacing the select value with null value.
- Calculating percentage of null values in each columns.
- Drop columns having null values more than 45%.
- Check the percentage of null values after dropping the columns having null values more than 45%.

Still there are many columns which contains null value . We will check one by one and impute if necessary or else will drop the column.

# DATA CLEANING (contd…..)

**Categorical Columns containing null values**

Dropping ‘CITY’ column

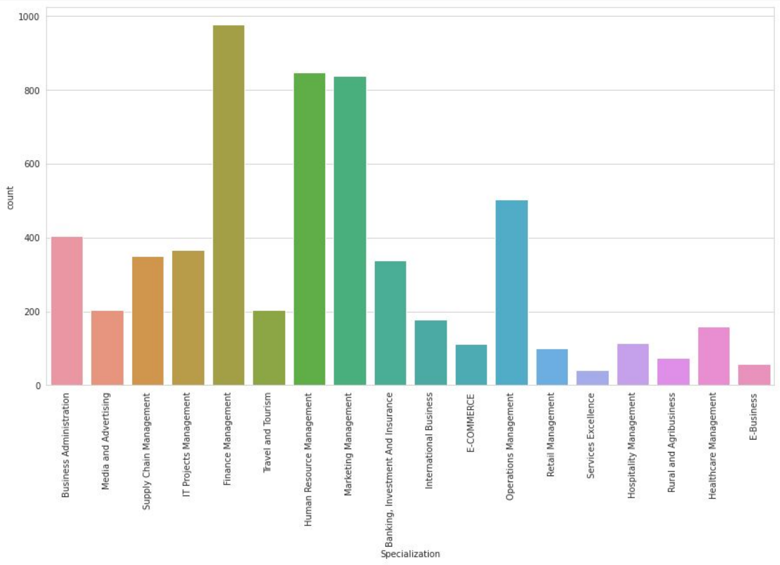



As the data have almost 40% of missing values . If we try to impute it then it will be skewed towards a particular value. Moreover it is a online platform and hence it can be access online so if we drop the column city it will not affect our analysis much.Therefore we can drop ‘city’ column.

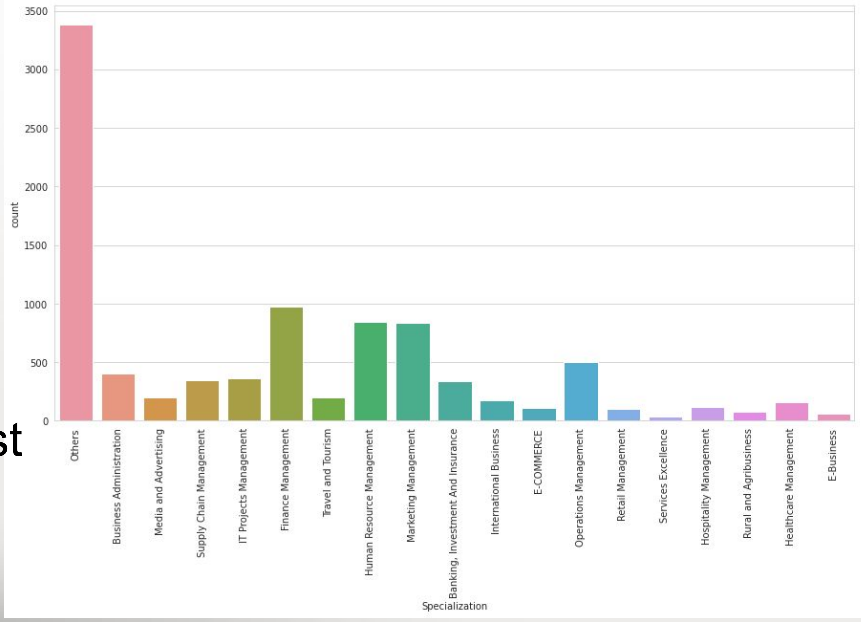
# DATA CLEANING (contd…..)

**Imputing null values of Specialization columns**

First of all we will calculate null value % in specialization columns and will find out percentage of unique values.




We will impute the null value with Others as it may be possible that the user may have no work experience or is a student.

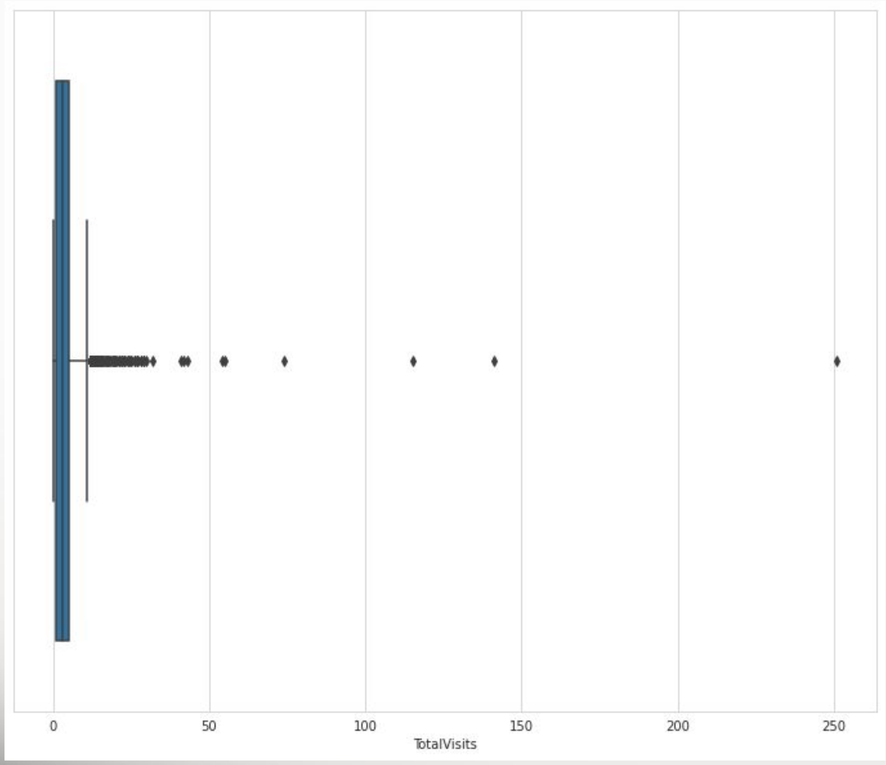



Similarly we will impute null values of columns such as Tags, What matters most to you in choosing a course, What is your current occupation, country, Last activity, Lead source.

# DATA CLEANING (contd…..)

**Numericals columns containing null values.**

Checking the null values , statistical part  and outliers  in Total visits.

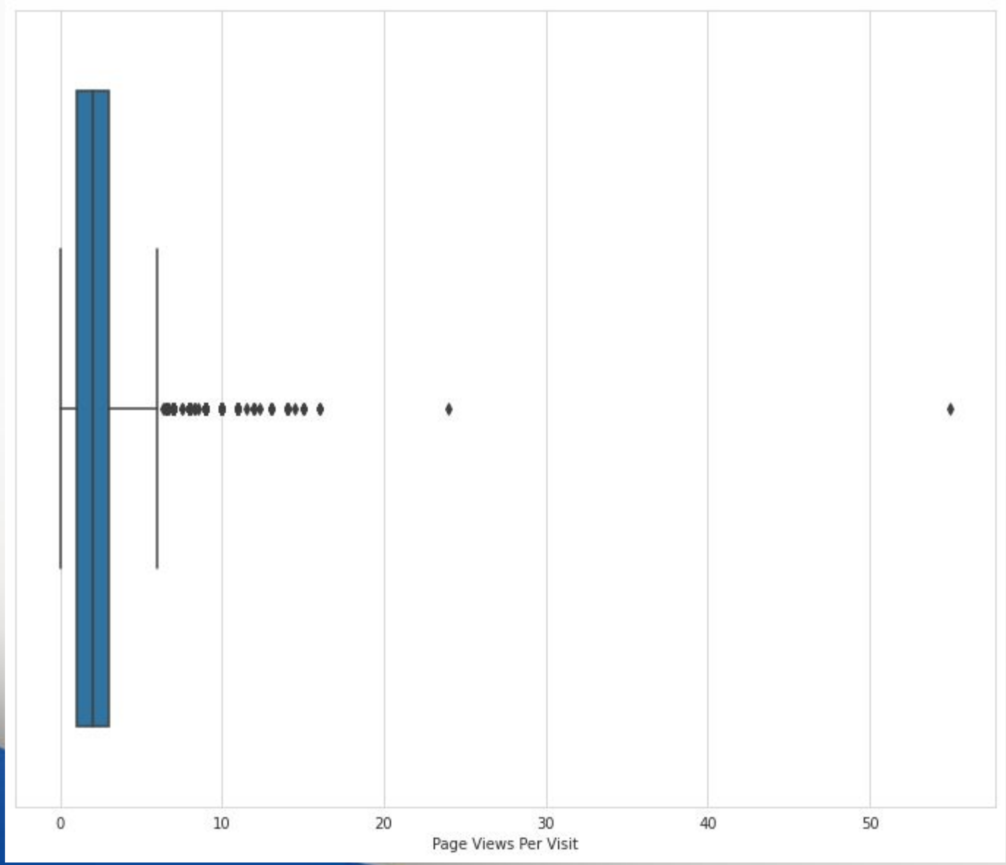



Since  the column contains outliers therefore we will impute null values with median

# DATA READING AND CLEANING (contd…..)

**Numericals columns containing null values.**

Checking the null values , statistical part  and outliers  in Page Views Per Visit.

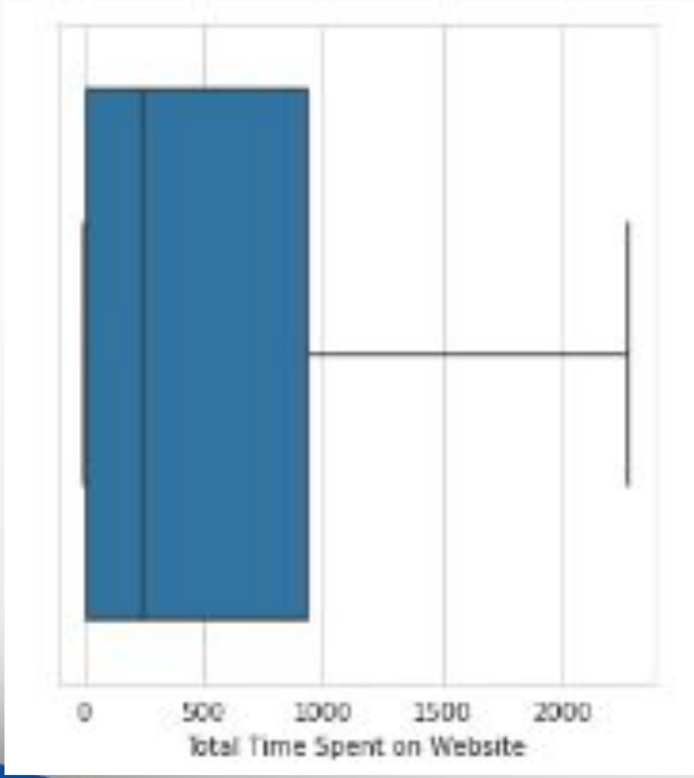



Since  the column contains outliers therefore we will impute null values with median

# DATA READING AND CLEANING (contd…..)

**Numericals columns containing null values.**

Checking the null values , statistical part  and outliers  in Total time spent on website.

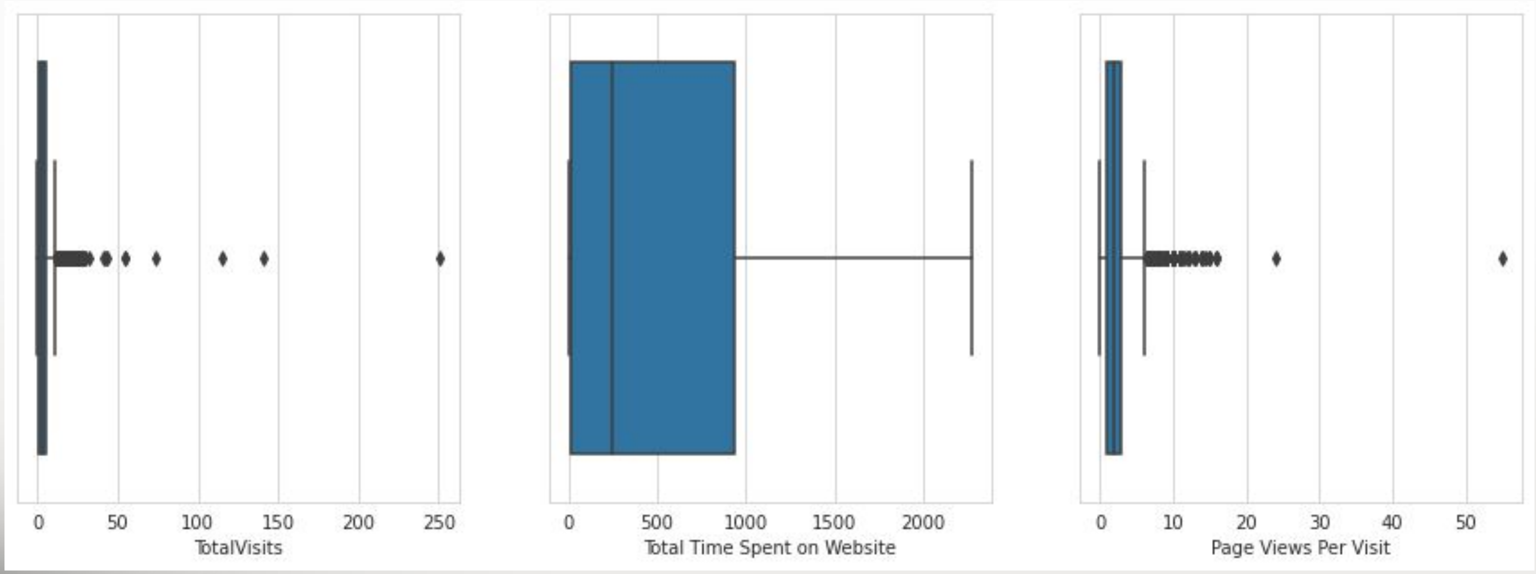



Time on Website is free from outliers

# DATA CLEANING (contd…..)

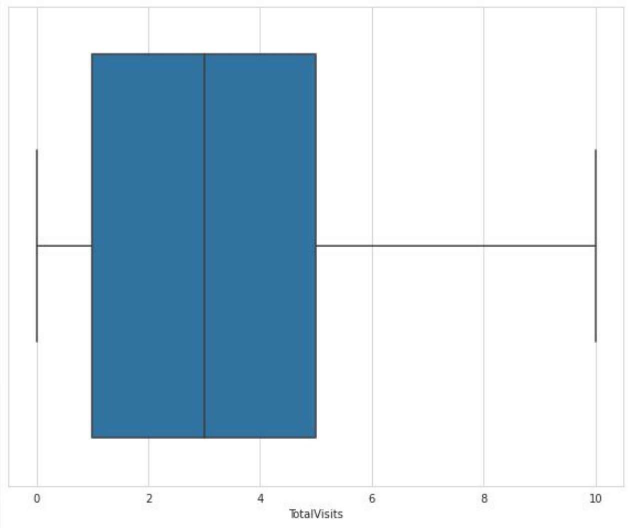
**Outliers Treatment**

From below given graphs we can see that Total Visits and Page views Per Visit columns contains outliers hence need to be taken care. While column Time on Website is free from outliers

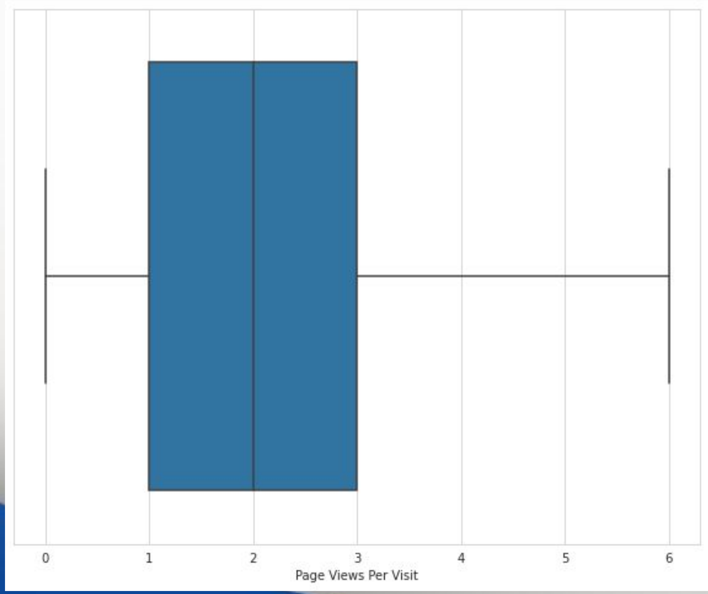


# DATA READING AND CLEANING (contd…..)

- As there are outliers in both the columns but outliers are valid value so we will cap them. To retain the data . 99% data will be capped to 95% as they very close hence impact will be same.



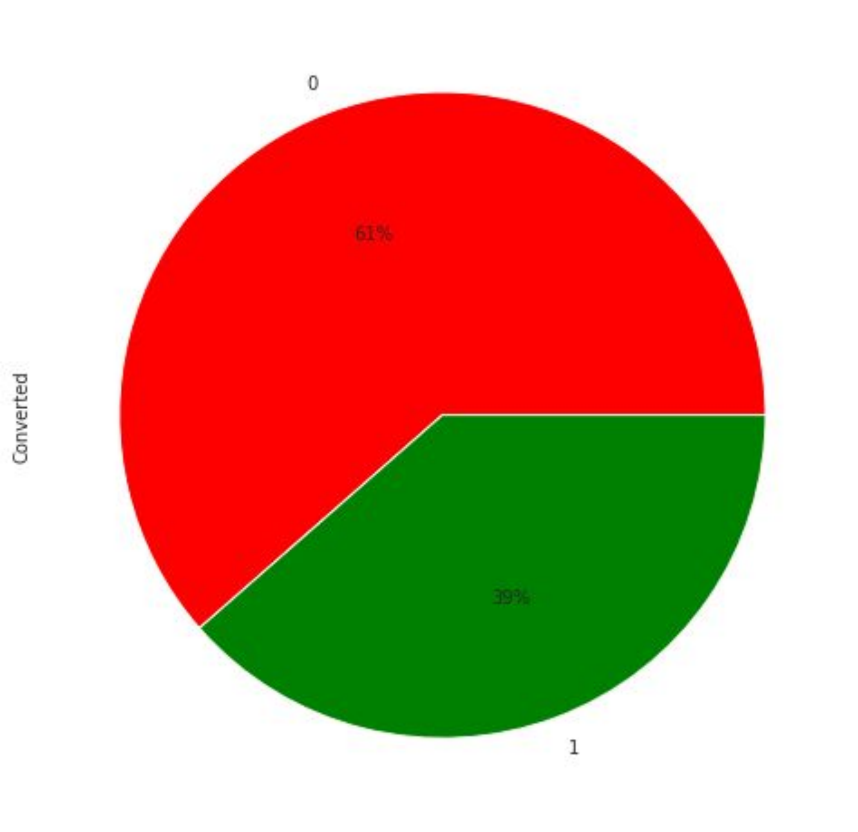

Box plot Graph to view quantile view after removal of outliers in Total visits columns




Box plot Graph to view quantile view after removal of outliers in Pages views per visits columns

# EXPLORATORY DATA  ANALYSIS (contd…..)

- Univarite and Bivariate Analysis on Categorical Variables




- Percentage of data imbalance

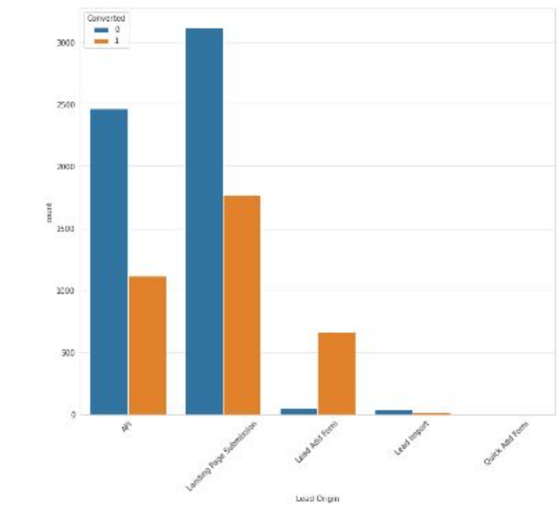
From this pie chart we can see that converted values are 39% and non converted values are 61%

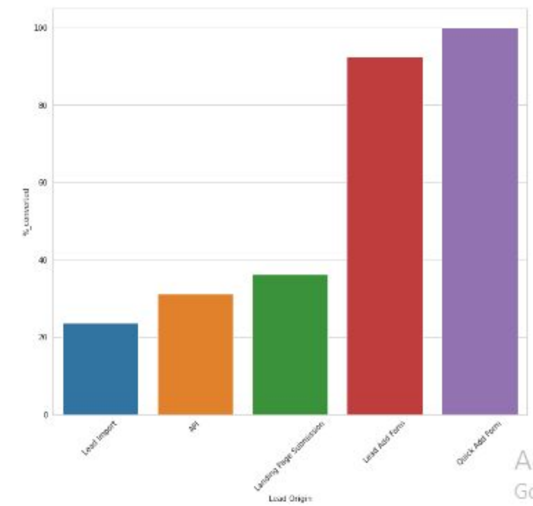
# EXPLORATORY DATA ANALYSIS (contd…..)

- Univariate Analysis

- Checking value counts of Lead Origin column Lead Origin.
- visualizing count of Lead Origin based on Converted value

| | Lead Origin | Lead_converted | Total_Count | %_converted |
|---|---|---|---|---|
| 0 | Lead Import | 13 | 55 | 23.64 |
| 1 | API | 1115 | 3580 | 31.15 |
| 2 | Landing Page Submission | 1768 | 4886 | 36.19 |
| 3 | Lead Add Form | 664 | 718 | 92.48 |
| 4 | Quick Add Form | 1 | 1 | 100.00 |



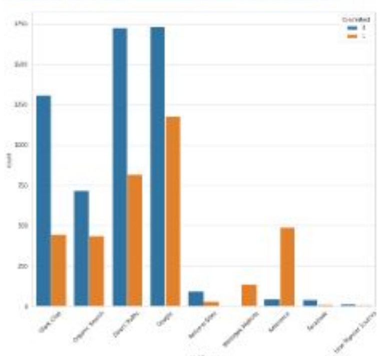



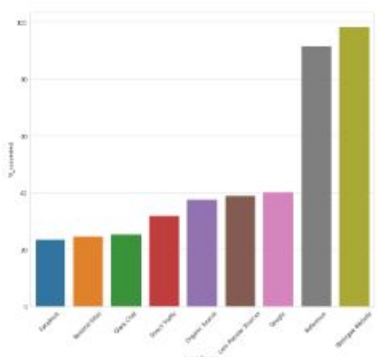
# EXPLORATORY DATA ANALYSIS (contd…..)

- Univariate Analysis

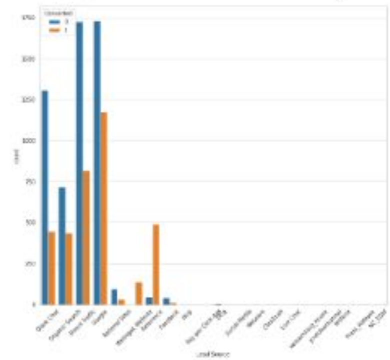
- Checking value counts of Lead Origin column Lead Source.
- visualizing count of Lead source based on Converted value

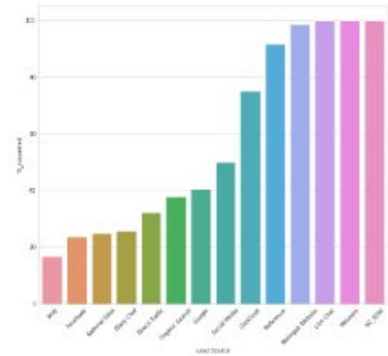
| | Lead Source | Lead converted | Total Count | % converted |
|---|---|---|---|---|
| 0 | Facebook | 13 | 55 | 23.64 |
| 1 | Referral Sites | 31 | 125 | 24.80 |
| 2 | Olark Chat | 448 | 1755 | 25.53 |
| 3 | Direct Traffic | 818 | 2543 | 32.17 |
| 4 | Organic Search | 436 | 1154 | 37.78 |
| 5 | Less Popular Sources | 9 | 23 | 39.13 |
| 6 | Google | 1176 | 2909 | 40.43 |
| 7 | Reference | 490 | 534 | 91.76 |
| 8 | Welingak Website | 140 | 142 | 98.59 |





| | Lead Source | Lead converted | Total Count | % converted |
|---|---|---|---|---|
| 0 | bing | 1 | 6 | 16.67 |
| 1 | Facebook | 13 | 55 | 23.64 |
| 2 | Referral Sites | 31 | 125 | 24.80 |
| 3 | Olark Chat | 448 | 1755 | 25.53 |
| 4 | Direct Traffic | 818 | 2543 | 32.17 |
| 5 | Organic Search | 436 | 1154 | 37.78 |
| 6 | Google | 1176 | 2909 | 40.43 |
| 7 | Social Media | 1 | 2 | 50.00 |
| 8 | Click2call | 3 | 4 | 75.00 |
| 9 | Reference | 490 | 534 | 91.76 |
| 10 | Welingak Website | 140 | 142 | 98.59 |
| 11 | Live Chat | 2 | 2 | 100.00 |
| 12 | WeLearn | 1 | 1 | 100.00 |
| 13 | NC_EDM | 1 | 1 | 100.00 |





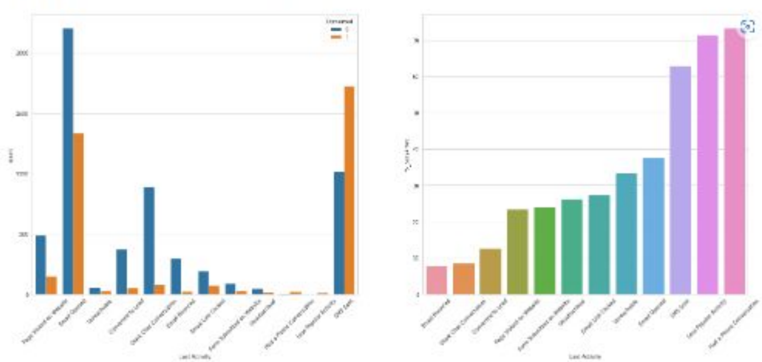
- Visualizing count lead source after merging nearby % value of converted value

# EXPLORATORY DATA ANALYSIS (contd…..)

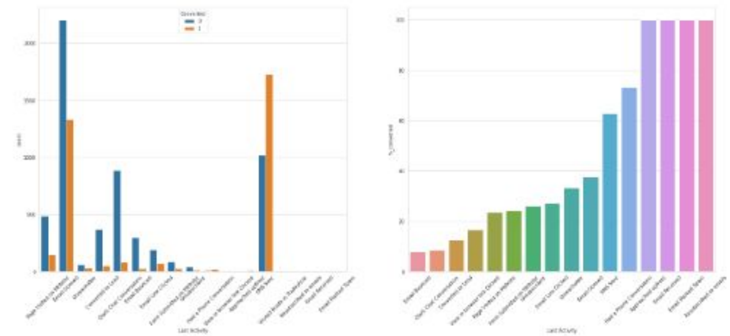
- Univariate Analysis

- Checking value counts of Lead Origin column Last activity.
- visualizing count of Last activity based on Converted value

| | Last Activity | Lead converted | Total Count | % converted |
|---|---|---|---|---|
| 0 | Email Bounced | 26 | 326 | 7.98 |
| 1 | Olark Chat Conversation | 84 | 973 | 8.63 |
| 2 | Converted to Lead | 54 | 428 | 12.62 |
| 3 | Page Visited on Website | 151 | 640 | 23.59 |
| 4 | Form Submitted on Website | 28 | 116 | 24.14 |
| 5 | Unsubscribed | 16 | 61 | 26.23 |
| 6 | Email Link Clicked | 73 | 267 | 27.34 |
| 7 | Unreachable | 31 | 93 | 33.33 |
| 8 | Email Opened | 1334 | 3540 | 37.68 |
| 9 | SMS Sent | 1727 | 2745 | 62.91 |
| 10 | Less Popular Activity | 15 | 21 | 71.43 |
| 11 | Had a Phone Conversation | 22 | 30 | 73.33 |



| | Last Activity | Lead converted | Total Count | % converted |
|---|---|---|---|---|
| 0 | Email Bounced | 26 | 326 | 7.98 |
| 1 | Olark Chat Conversation | 84 | 973 | 8.63 |
| 2 | Converted to Lead | 54 | 428 | 12.62 |
| 3 | View in browser link Clicked | 1 | 6 | 16.67 |
| 4 | Page Visited on Website | 151 | 640 | 23.59 |
| 5 | Form Submitted on Website | 28 | 116 | 24.14 |
| 6 | Unsubscribed | 16 | 61 | 26.23 |
| 7 | Email Link Clicked | 73 | 267 | 27.34 |
| 8 | Unreachable | 31 | 93 | 33.33 |
| 9 | Email Opened | 1334 | 3540 | 37.68 |
| 10 | SMS Sent | 1727 | 2745 | 62.91 |
| 11 | Had a Phone Conversation | 22 | 30 | 73.33 |
| 12 | Approached upfront | 9 | 9 | 100.00 |
| 13 | Email Received | 2 | 2 | 100.00 |
| 14 | Email Marked Spam | 2 | 2 | 100.00 |
| 15 | Resubscribed to emails | 1 | 1 | 100.00 |



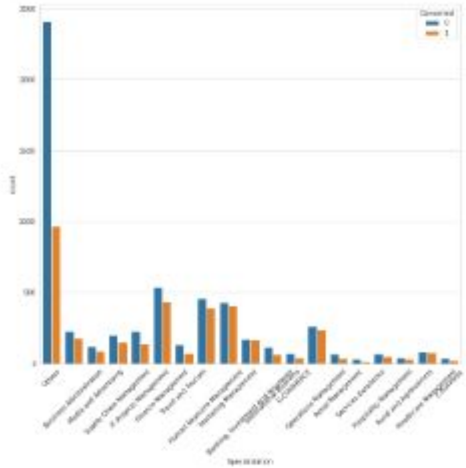
- Visualizing count lead source after merging nearby % value of converted value

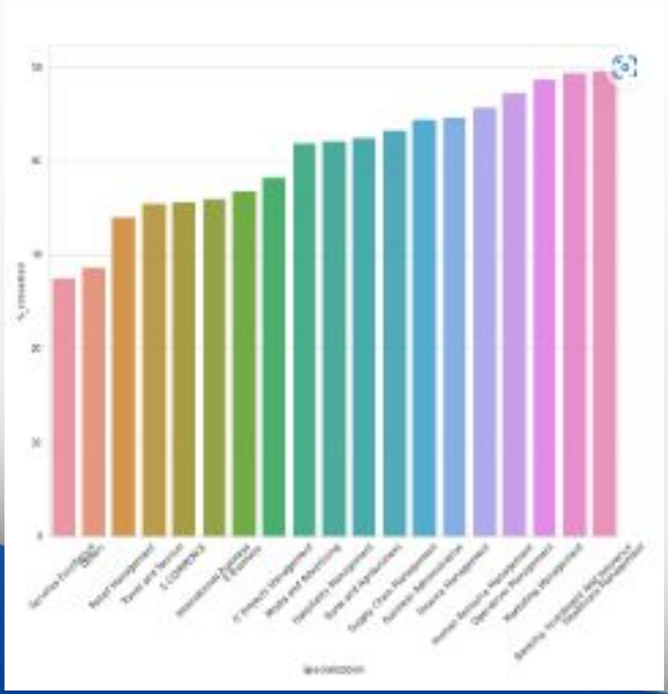
# EXPLORATORY DATA ANALYSIS (contd…..)

- Univariate Analysis

- Checking value counts of Lead Origin column Specialization.
- visualizing count of Specialization based on Converted value

|  | Specialization | Lead converted | Total Count | % converted |
|---|---|---|---|---|
| 0 | Services Excellence | 11 | 40 | 27.50 |
| 1 | Others | 969 | 3380 | 28.67 |
| 2 | Retail Management | 34 | 100 | 34.00 |
| 3 | Travel and Tourism | 72 | 203 | 35.47 |
| 4 | E-COMMERCE | 40 | 112 | 35.71 |
| 5 | International Business | 64 | 178 | 35.96 |
| 6 | E-Business | 21 | 57 | 36.84 |
| 7 | IT Projects Management | 140 | 366 | 38.25 |
| 8 | Media and Advertising | 85 | 203 | 41.87 |
| 9 | Hospitality Management | 48 | 114 | 42.11 |
| 10 | Rural and Agribusiness | 31 | 73 | 42.47 |
| 11 | Supply Chain Management | 151 | 349 | 43.27 |
| 12 | Business Administration | 179 | 403 | 44.42 |
| 13 | Finance Management | 436 | 976 | 44.67 |
| 14 | Human Resource Management | 388 | 848 | 45.75 |
| 15 | Operations Management | 238 | 503 | 47.32 |
| 16 | Marketing Management | 408 | 838 | 48.69 |
| 17 | Banking, Investment And Insurance | 167 | 338 | 49.41 |
| 18 | Healthcare Management | 79 | 159 | 49.69 |





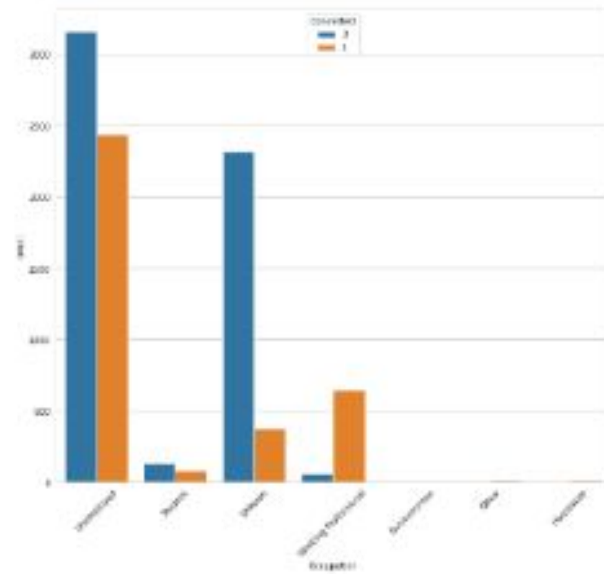
- Visualizing count lead source after merging nearby % value of converted value

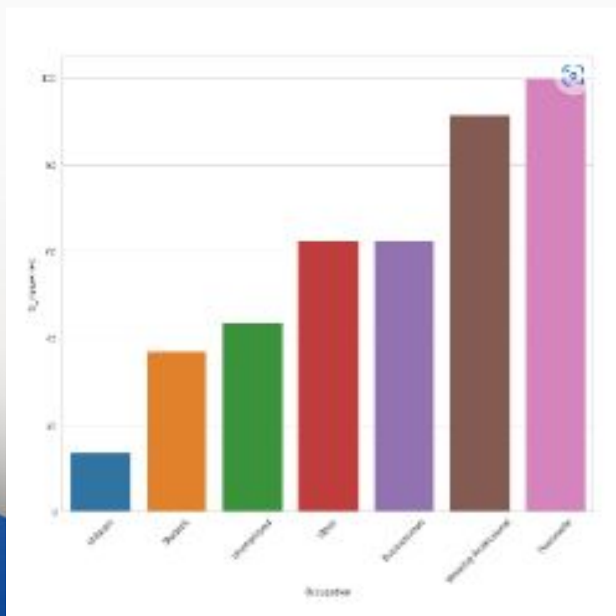
# EXPLORATORY DATA ANALYSIS (contd…..)

- Univariate Analysis

- Checking value counts of Lead Origin column Occupation.
- visualizing count of occupation based on Converted value

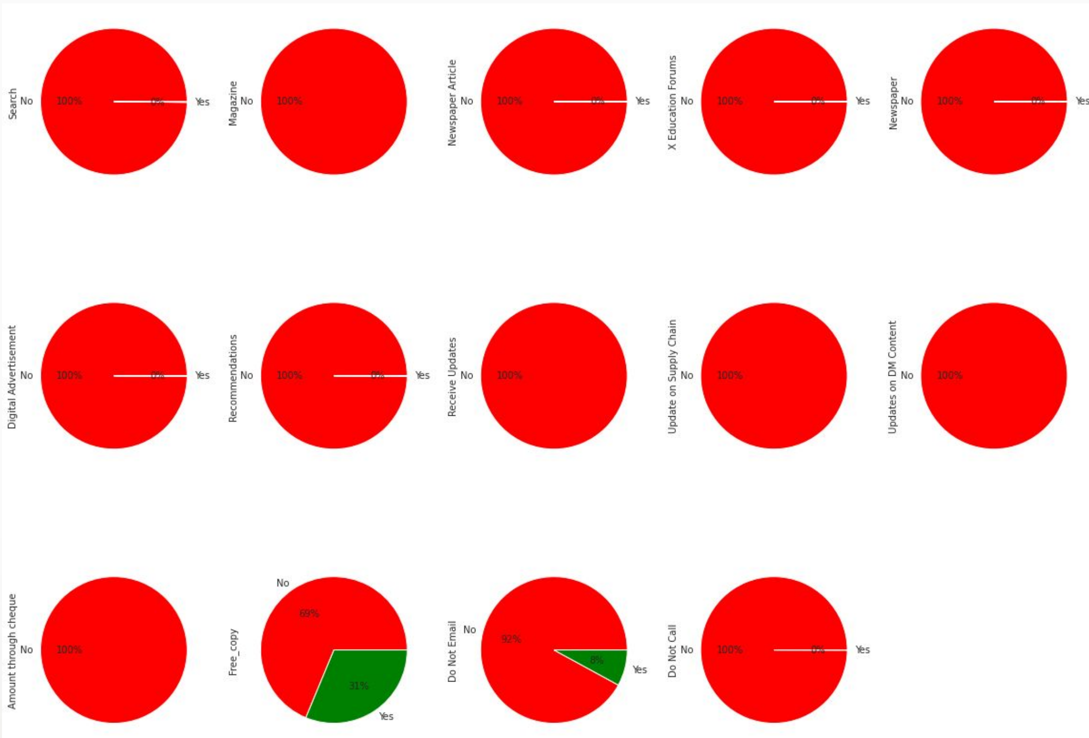
| | Occupation | Lead converted | Total Count | % converted |
|---|---|---|---|---|
| 0 | Unkown | 370 | 2690 | 13.75 |
| 1 | Student | 78 | 210 | 37.14 |
| 2 | Unemployed | 2441 | 5600 | 43.59 |
| 3 | Other | 10 | 16 | 62.50 |
| 4 | Businessman | 5 | 8 | 62.50 |
| 5 | Working Professional | 647 | 706 | 91.64 |
| 6 | Housewife | 10 | 10 | 100.00 |





- Visualizing count lead source after merging nearby % value of converted value

# EXPLORATORY DATA ANALYSIS (contd…..)

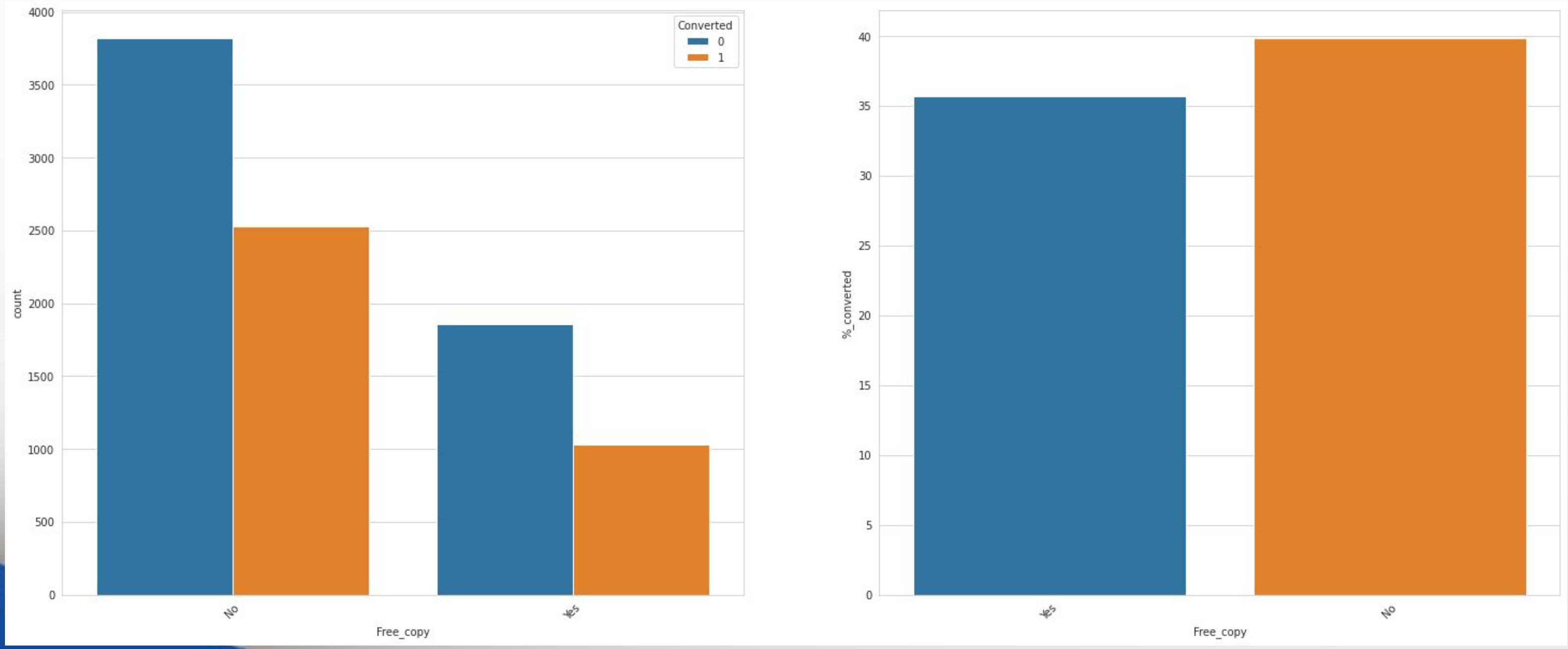



As shown in above piec harts it shows that Do not Email and Free copy has some positive values of conversion while others columns having 0% of conversion values . Therefore , we will analyze the columns with postive converiosn values while  rest of the column can be dropped as it is skewed towards one value

# EXPLORATORY DATA ANALYSIS (contd…..)

- Univariate analysis of ‘Free Copy columns’ .

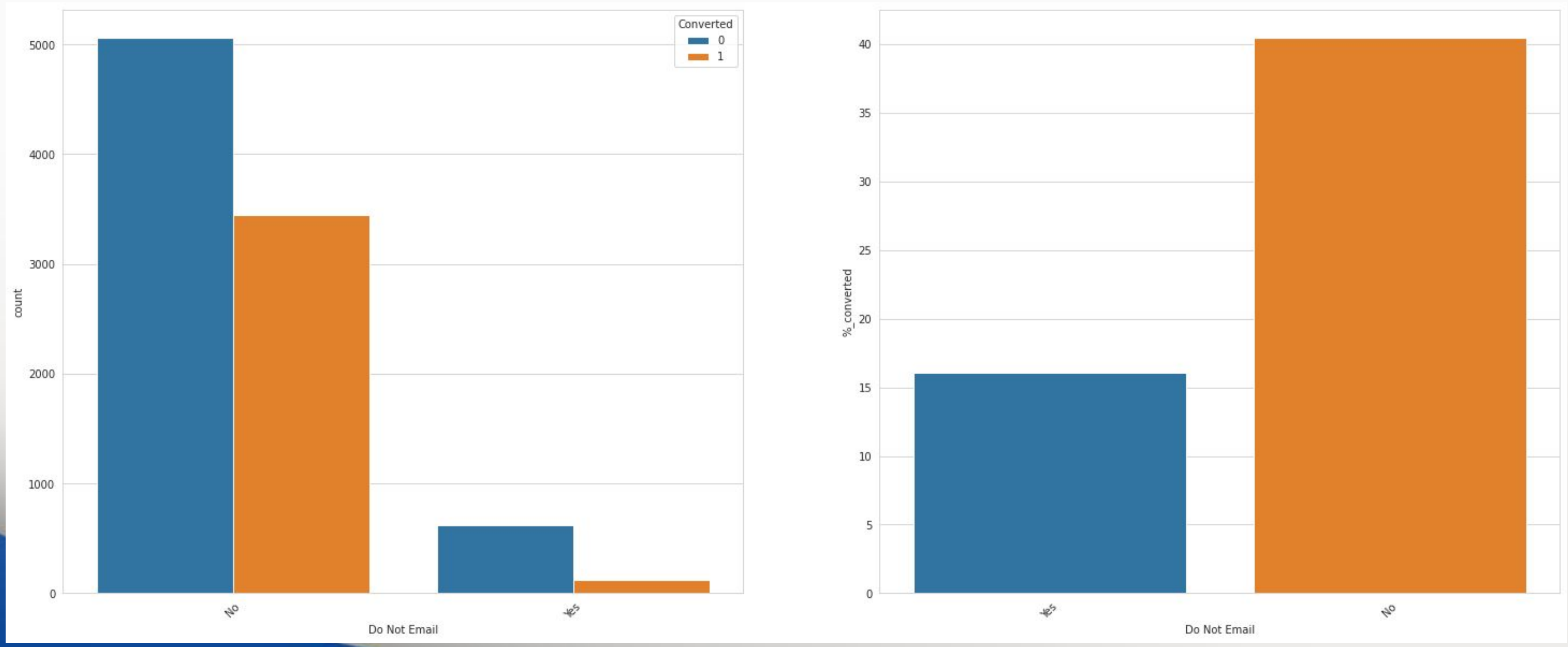
As the Convergence rate for both are almost same hence we need to drop these columns more over this columns doesn't add much value



# EXPLORATORY DATA ANALYSIS (contd…..)

- Univariate analysis of ‘Do not Email’ .

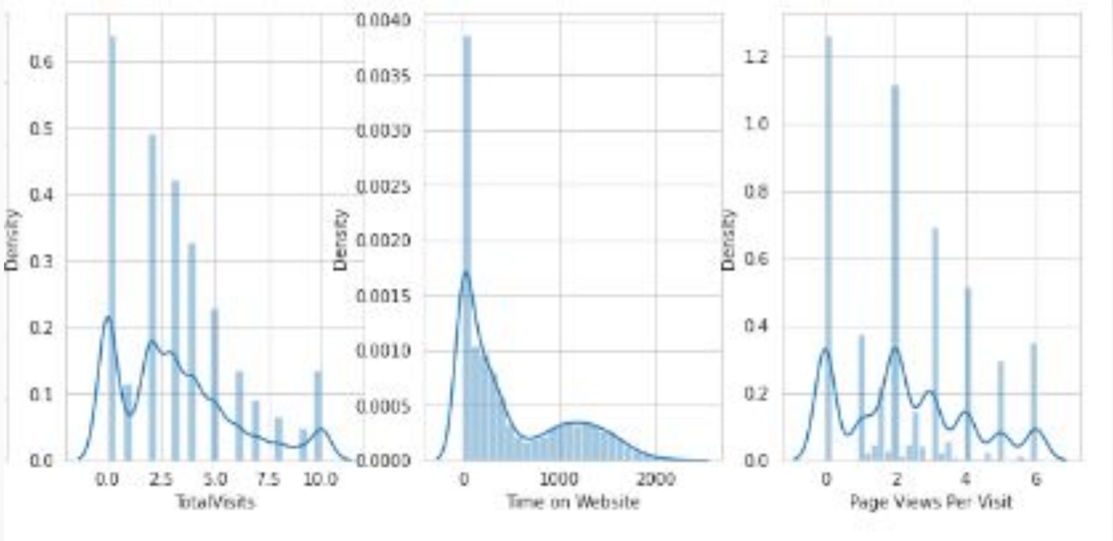
Maximum number of people prefer receiving mail. 40 percent of people are converted who prefer receiving mail.

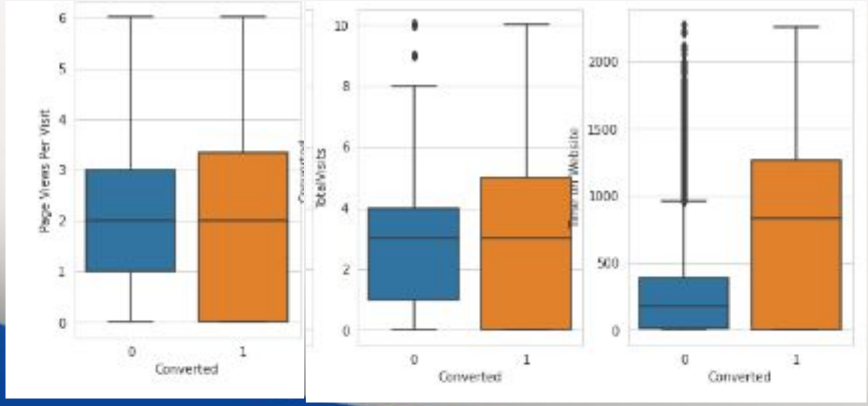


# EXPLORATORY DATA ANALYSIS (contd…..)

- Univariate analysis on numericals variable.

checking distribution of data in graphical presentation

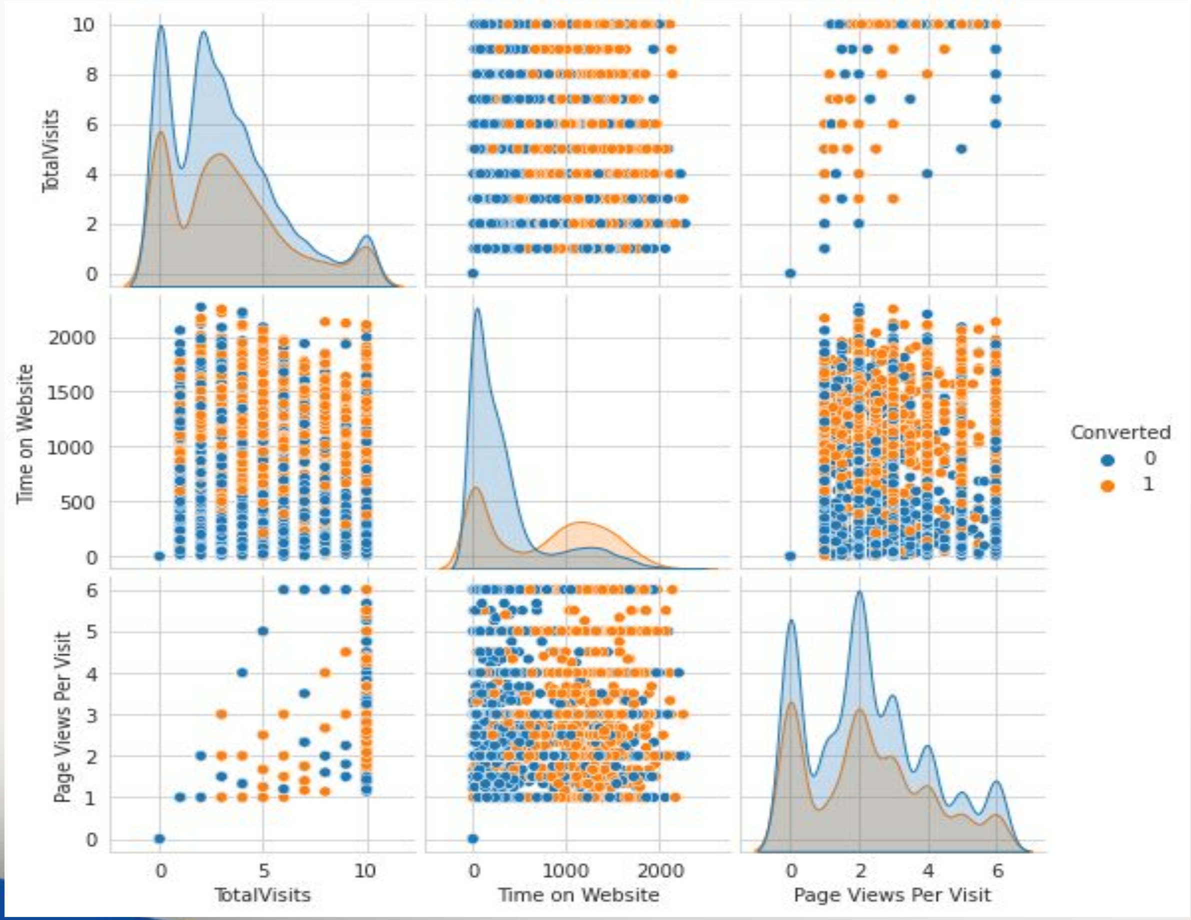



checking distribution of data in quantile range using boxplot

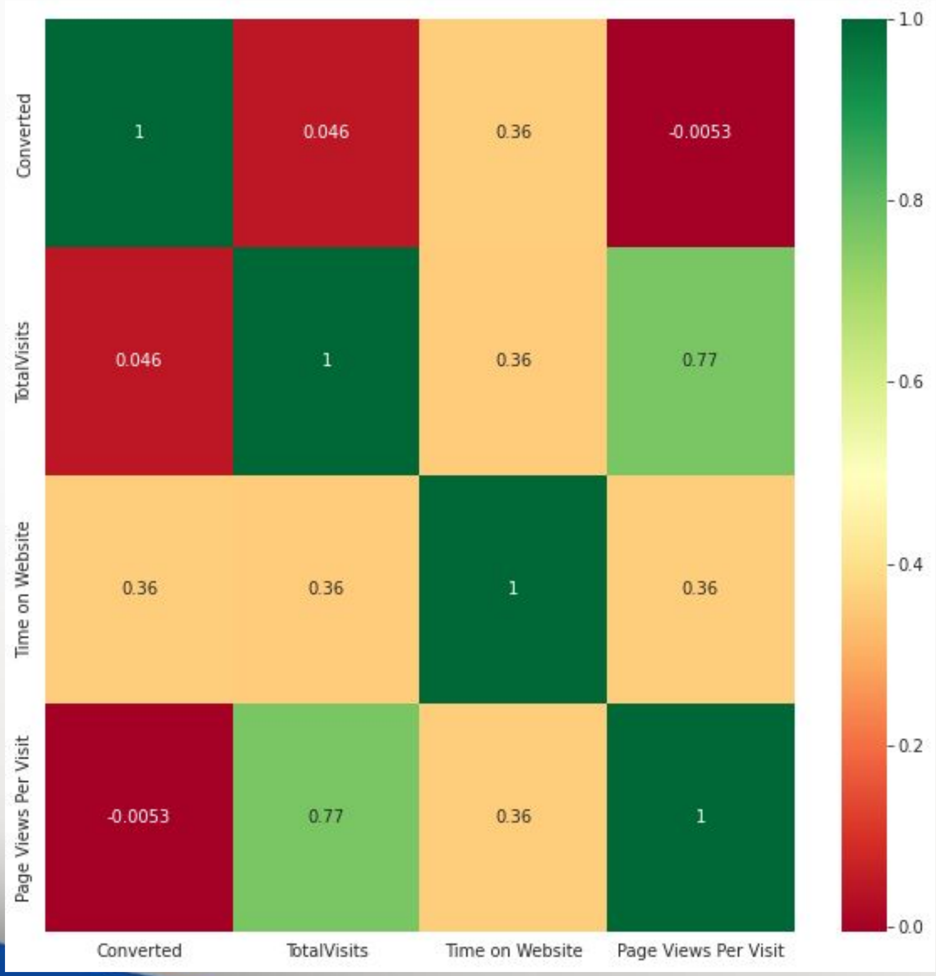
# EXPLORATORY DATA ANALYSIS (contd…..)

Bivariate analysis and checking for visible pattern



# EXPLORATORY DATA ANALYSIS (contd…..)

- Multivariate analysis through heat map



# Data Preparation for Model Building

- mapping Do Not Email column containing Yes and No to 1 and 0
- Do Not Email

| | |
|---|---|
| No | 8506 |
| Yes | 734 |

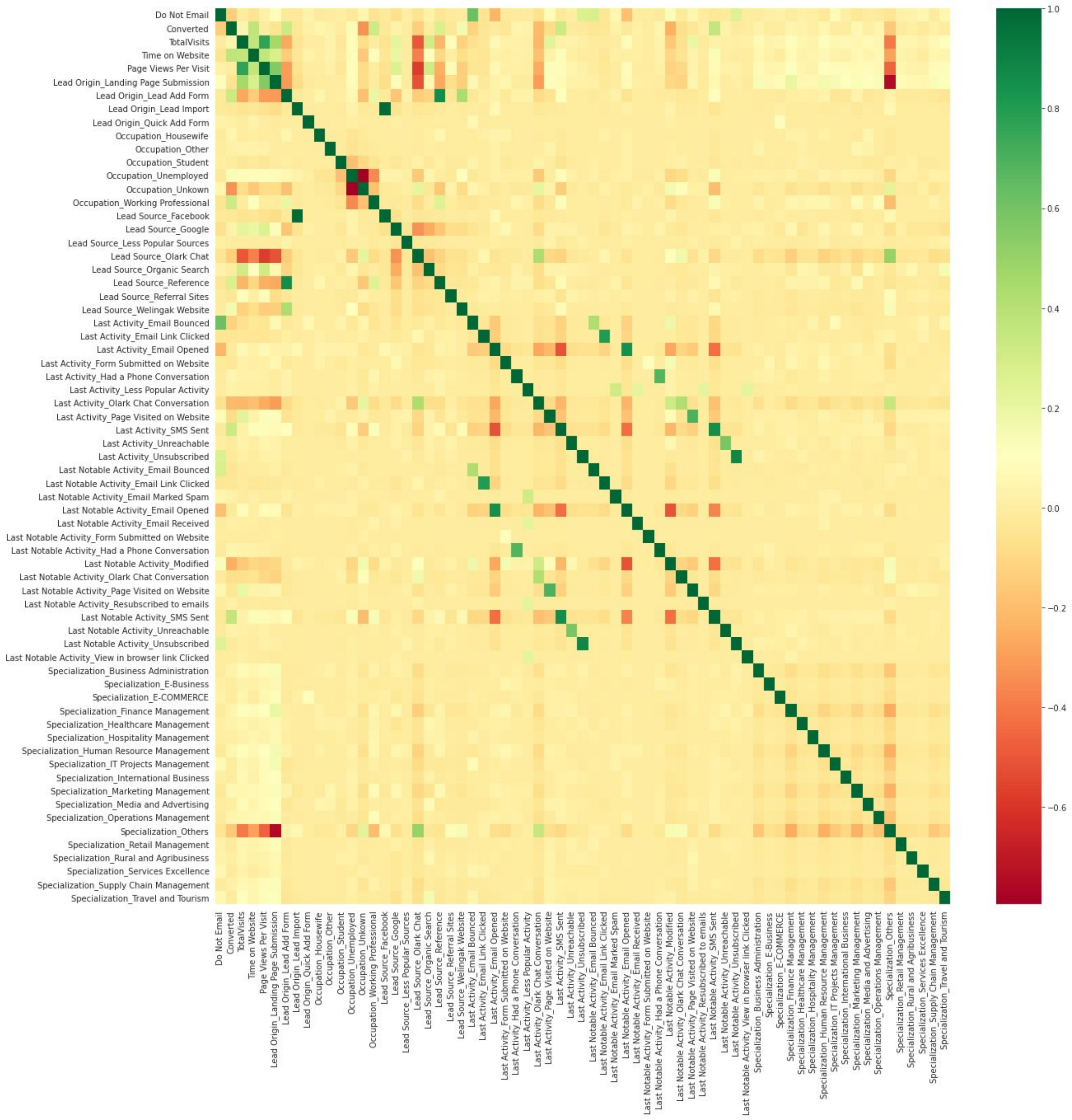
- Dummy variable creation of following columns :

✔ Lead Origin
✔ Occupation
✔ Lead Source
✔ Last Activity
✔ Last Notable Activity
✔ Specialization

| | Lead Origin | Lead Source | Do Not Email | Converted | TotalVisits | Time on Website | Page Views Per Visit | Last Activity | Specialization | Occupation | Last Notable Activity |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | API | Olark Chat | No | 0 | 0.0 | 0 | 0.0 | Page Visited on Website | Others | Unemployed | Modified |
| 1 | API | Organic Search | No | 0 | 5.0 | 674 | 2.5 | Email Opened | Others | Unemployed | Email Opened |
| 2 | Landing Page Submission | Direct Traffic | No | 1 | 2.0 | 1532 | 2.0 | Email Opened | Business Administration | Student | Email Opened |
| 3 | Landing Page Submission | Direct Traffic | No | 0 | 1.0 | 305 | 1.0 | Unreachable | Media and Advertising | Unemployed | Modified |
| 4 | Landing Page Submission | Google | No | 1 | 2.0 | 1428 | 1.0 | Converted to Lead | Others | Unemployed | Modified |

Dataframe has total 9240 rows and 67 columns not required columns can be drop.

# Data Preparation for Model Building



# Logistic Regression Model Building

**Feature Scaling using Standard scaler**

- Scaling helps in interpretation. It is important to have all variables(specially categorical ones which has values 0 and 1) on the same scale for the model to be easily interpretable.

- Standardisation' was used to scale the data for modelling. It basically brings all of the data into a standard normal distribution with mean at zero and standard deviation one.

| | TotalVisits | Time on Website | Page Views Per Visit |
|---|---|---|---|
| **1871** | -1.149699 | -0.885371 | -1.266675 |
| **6795** | 0.299722 | 0.005716 | -0.516439 |
| **3516** | 0.662077 | -0.691418 | 0.143543 |
| **8105** | 0.662077 | 1.365219 | 1.553761 |
| **3934** | -1.149699 | -0.885371 | -1.266675 |

# Logistic Regression Model Building (CONTD)

## Feature Selection Using RFE

- Recursive feature elimination is an optimization technique for finding the best performing subset of features. It is based on the idea of repeatedly constructing a model and choosing either the best (based on coefficients), setting the feature aside and then repeating the process with the rest of the features. This process is applied until all the features in the dataset are exhausted. Features are then ranked according to when they were eliminated.

```
# recursive featur elemination
log_reg = LogisticRegression()

# running RFE with 20 variables
rfe = RFE(log_reg,n_features_to_select= 20)

rfe=rfe.fit(X_train ,y_train)
```

Running RFE with the output number of the variable equal to 20

```
col=list(X_train.columns[rfe.support_])
col
```

```
['Do Not Email',
 'Time on Website',
 'Lead Origin_Landing Page Submission',
 'Lead Origin_Lead Add Form',
 'Occupation_Housewife',
 'Occupation_Unkown',
 'Occupation_Working Professional',
 'Lead Source_Olark Chat',
 'Lead Source_Welingak Website',
 'Last Activity_Email Opened',
 'Last Activity_Had a Phone Conversation',
 'Last Activity_Less Popular Activity',
 'Last Activity_SMS Sent',
 'Last Activity_Unsubscribed',
 'Last Notable Activity_Had a Phone Conversation',
 'Last Notable Activity_Modified',
 'Last Notable Activity_Olark Chat Conversation',
 'Last Notable Activity_Unreachable',
 'Specialization_Hospitality Management',
 'Specialization_Others']
```

# Logistic Regression Model Building (CONTD)

- Manual Feature selection for different Models using below given steps :

- Generalized model is built initially with the 18 variables selected by RFE.

- Unwanted features are dropped serially after checking p values (< 0.5) and VIF (< 5) and model is built multiple times.

- The final model with 14 features, passes both the significance test and the multicollinearity test.

# Logistic Regression Model Building (CONTD)

**Model Evaluation**

**Predicting the value on train set.**

Predicating the conversion probability and the predicted columns

Creating a dataframe with Actual Predicted and Predicted Probabilities

| | Converted | Converted_Probability | ID |
|---|---|---|---|
| 0 | 0 | 0.257218 | 1871 |
| 1 | 0 | 0.232387 | 6795 |
| 2 | 0 | 0.300304 | 3516 |
| 3 | 0 | 0.809566 | 8105 |
| 4 | 0 | 0.132476 | 3934 |

Creating new column 'predicted‘ 1 if probability is greater than 0.5 or 0 if it is less than 0.5

Showing top 5 records of the dataframe in the picture on the left.

| | Converted | Converted_Probability | ID | Predict |
|---|---|---|---|---|
| 0 | 0 | 0.257218 | 1871 | 0 |
| 1 | 0 | 0.232387 | 6795 | 0 |
| 2 | 0 | 0.300304 | 3516 | 0 |
| 3 | 0 | 0.809566 | 8105 | 1 |
| 4 | 0 | 0.132476 | 3934 | 0 |

Confusion Matrix

```
[[3564  438]
 [ 734 1732]]
```

# Logistic Regression Model Building (CONTD)

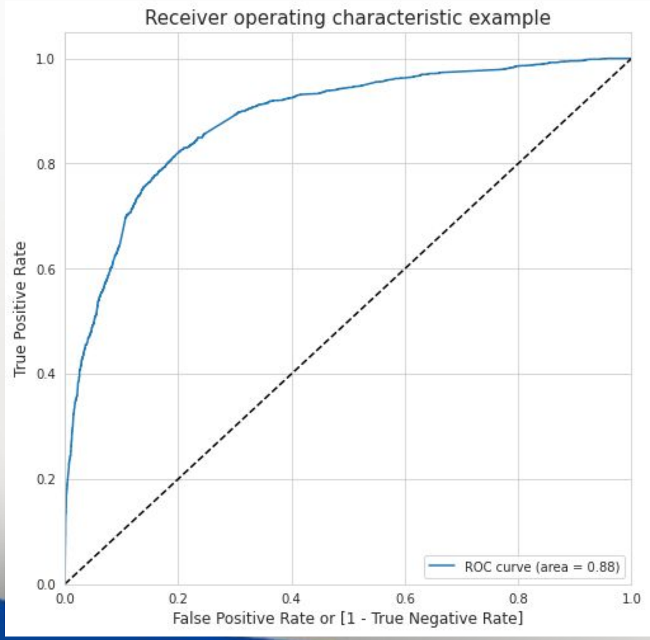
**Metrics Will be used for Evaluation** - Accuracy ,Sensitivity, Specificity, Precision, Recall ,True Positive Rate , True Negative Rate,False Positive Rate ,False Negative Rate, Postitive Predictive Value and Negative Predictive Value

| Predicted | not_converted | converted |
|---|---|---|
| Actual | | |
| not_converted | 3564 | 438 |
| converted | 734 | 1732 |

```
Model Accuracy value is                 =  81.88 %
Model Sensitivity value is              =  70.24 %
Model Specificity value is              =  89.06 %
Model Precision value is                =  79.82 %
Model Recall value is                   =  70.24 %
Model True Positive Rate                =  70.24 %
Model False Positive Rate               =  10.94 %
Model Poitive Prediction Value is       =  79.82 %
Model Negative Prediction value is      =  82.92 %
```

# Receiver Operating Characteristics (ROC) Curve

- **Plotting ROC Curve based on Training set data**
- Benefits of ROC Curve:-
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- 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



# Optimal Cutoff Point at Training set

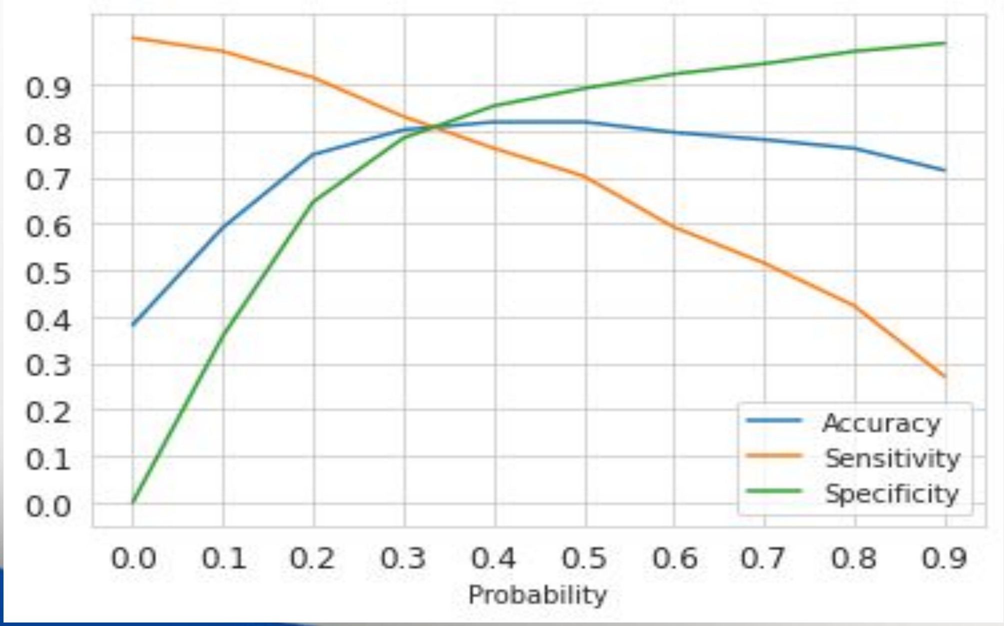
- Optimal cutoff probability is that prob where we get balanced sensitivity and specificity.
- TP = confusion[1,1] # true positive
- TN = confusion[0,0] # true negatives
- FP = confusion[0,1] # false positives
- FN = confusion[1,0] # false negatives

- **Sensitivity** with respect to our model can be defined as the ratio of total number of actual Conversions correctly predicted to the total no of actual Conversions.

- Similarly, **Specificity** can be defined as the ratio of total no of actual non-Conversions correctly predicted to the total number of actual non-Conversions.

# Optimal Cutoff Point at Training set

- At first we have randomly taken 0.5 as our cut-off point now we will use the Optimal cut-off point to determine the cut-off value and calculate the Evaluation Metrics once again.
- The cut off point is 0.34 .



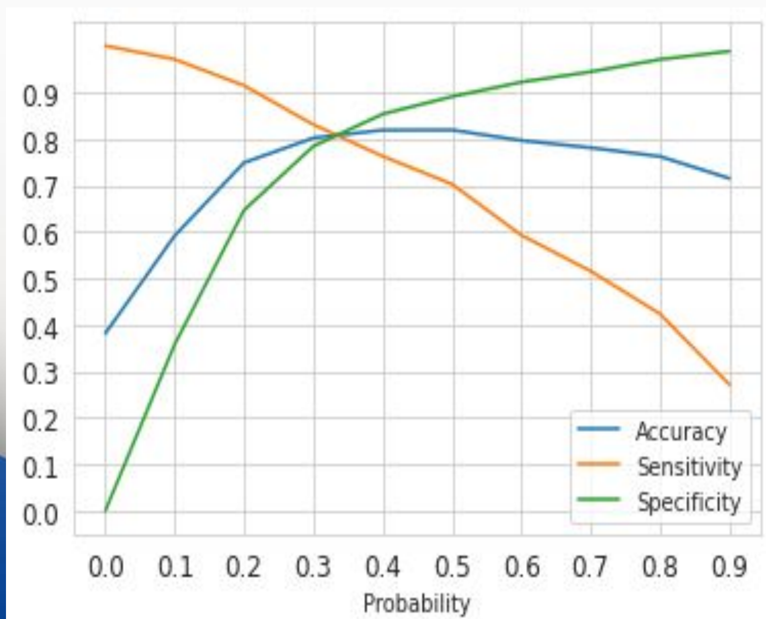
# Observation of Training set

- We have the following values for the Train Data:
- --> Accuracy  : 81.11%
- --> Sensitivity : 80.29 %
- --> Specificity : 81.61%
- → F1 score   : 0.7641

| | | |
|---|---|---|
| Model Accuracy value is | = | 81.11 % |
| Model Sensitivity value is | = | 80.29 % |
| Model Specificity value is | = | 81.61 % |
| Model Precision value is | = | 72.9 % |
| Model Recall value is | = | 80.29 % |
| Model True Positive Rate | = | 80.29 % |
| Model False Positive Rate | = | 18.39 % |
| Model Poitive Prediction Value is | = | 72.9 % |
| Model Negative Prediction value is | = | 87.05 % |

# Optimal Probability Threshold at Training set

- plot accuracy sensitivity and specificity for various probabilities.
- The accuracy sensitivity and specificity was calculated for various values of probability threshold and plotted in the graph to the right. • From the curve above, 0.34 is found to be the optimum point for cutoff probability.
- At this threshold value, all the 3 metrics - accuracy sensitivity and specificity was found to be well above 80% which is a well acceptable value.




| | | |
|---|---|---|
| Model Accuracy value is | = | 81.11 % |
| Model Sensitivity value is | = | 80.29 % |
| Model Specificity value is | = | 81.61 % |
| Model Precision value is | = | 72.9 % |
| Model Recall value is | = | 80.29 % |
| Model True Positive Rate | = | 80.29 % |
| Model False Positive Rate | = | 18.39 % |
| Model Poitive Prediction Value is | = | 72.9 % |
| Model Negative Prediction value is | = | 87.05 % |

# Precision and Recall trade off

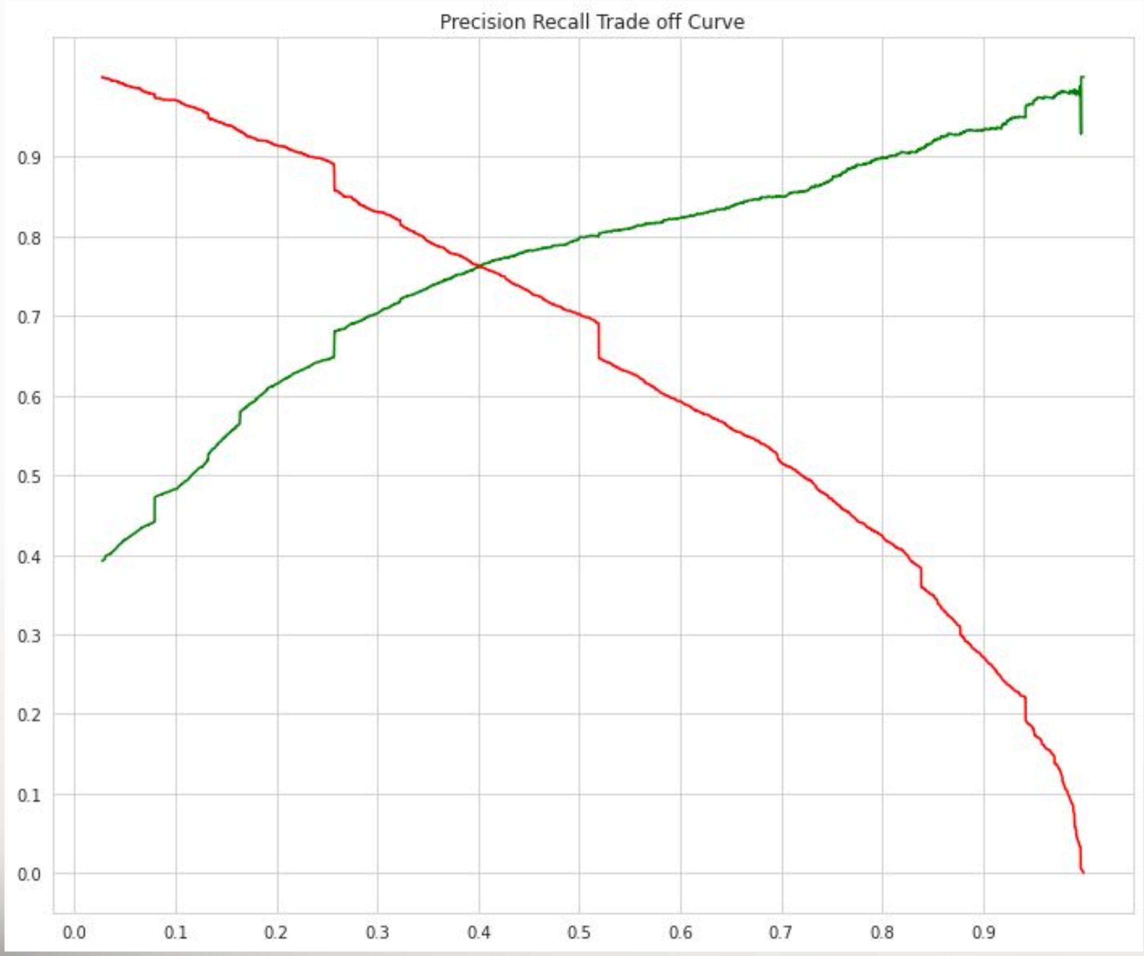
**Precision**

- Sensitivity TP / (TP+FN) : 0.729

**Recall**

- TP / TP + FN : 0.8029
- The final model on the train dataset is used to make predictions for the test dataset
- The train data set was scaled using the scaler. transform function that was used to scale the train dataset.
- The Predicted probabilities were added to the leads in the test dataframe.
- Using the probability threshold value of 0.34, the leads from the test dataset were predicted if they will convert or not.

# Precision and Recall trade off curve



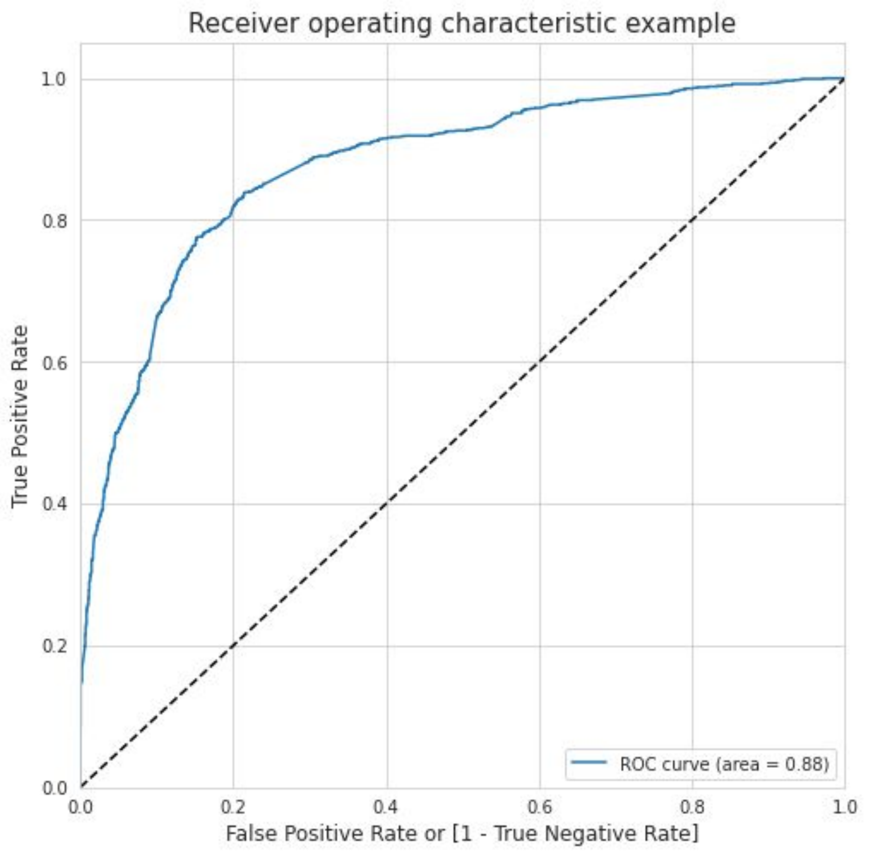
# ROC curve for Test set

**Making Prediction on the test set**

Observation:

ROC curve based on Test set :

- ROC curve value is 0.88 which shows that it is performing well on test set



# Observation of Test set

**Making Prediction on the test set and taking a cutoff of 0.34**

- We have the following values for the Test Data:
- --> Accuracy : 80.84%
- --> Sensitivity : 79.27 %
- --> Specificity : 81.87%
- —> F1 Score:-  0.7657

```
Model Accuracy value is                 =  80.84 %
Model Sensitivity value is              =  79.27 %
Model Specificity value is              =  81.87 %
Model Precision value is                =  74.06 %
Model Recall value is                   =  79.27 %
Model True Positive Rate                =  79.27 %
Model False Positive Rate               =  18.13 %
Model Poitive Prediction Value is       =  74.06 %
Model Negative Prediction value is      =  85.81 %
```

# Lead Score Calculation

Lead Score is calculated for all the leads in the original dataframe.

Formula for Lead Score calculation is :

Lead Score = 100 * Conversion Probability

| | ID | Converted | Converted_Probability | Final_Predicted | Lead_Score |
|---|---|---|---|---|---|
| 0 | 4269 | 1 | 0.779881 | 1 | 77 |
| 1 | 2376 | 1 | 0.941884 | 1 | 94 |
| 2 | 7766 | 1 | 0.929822 | 1 | 92 |
| 3 | 9199 | 0 | 0.079554 | 0 | 7 |
| 4 | 4359 | 1 | 0.838590 | 1 | 83 |
| 5 | 9186 | 1 | 0.548813 | 1 | 54 |
| 6 | 1631 | 1 | 0.467584 | 1 | 46 |
| 7 | 8963 | 1 | 0.201474 | 0 | 20 |
| 8 | 8007 | 0 | 0.053536 | 0 | 5 |
| 9 | 5324 | 1 | 0.327542 | 0 | 32 |

# Lead Score Calculation

•The train and test dataset is concatenated to get the entire list of leads available.

•The Conversion Probability is multiplied by 100 to obtain the Lead Score for each lead.

•Higher the lead score, higher is the probability of a lead getting converted and vice versa.

•Since, we had used 0.33 as our final Probability threshold for deciding if a lead will convert or not, any lead with a lead score of 34 or above will have a value of '1' in the final predicted column.

Final Observation:

Train Data:

Accuracy : 81.88%

Sensitivity : 80.29%

Specificity : 81.61%

Test Data:

Accuracy : 80.84%

Sensitivity : 79.27%

Specificity : 81.87%

The model seems to be performing well. Can be recommend this model in making good calls based on this model

# Determining Feature Importance

Selecting the coefficients of the selected features from our final model excluding the intercept.

•14 features have been used by our model to successfully predict if a lead will get converted or not.

•The Coefficient (beta) values for each of these features from the model parameters are used to determine the order of importance of these features.

•Features with high positive beta values are the ones that contribute most towards the probability of a lead getting converted.

•Similarly, features with high negative beta values contribute the least.

# Determining Feature Importance

- The Relative Importance of each feature is determined on a scale of 100 with the feature with having a score of 100.

```
top_features = 100.0 * (top_features / top_features.max())
```

Positive feature

| | |
|---|---|
| Last Notable Activity_Had a Phone Conversation | 100.000000 |
| Lead Origin_Lead Add Form | 90.834614 |
| Occupation_Working Professional | 72.829897 |
| Last Notable Activity_Unreachable | 53.291834 |
| Last Activity_Less Popular Activity | 52.521730 |
| Lead Source_Welingak Website | 47.066473 |
| Last Activity_SMS Sent | 43.458806 |
| Time on Website | 28.175552 |
| Lead Source_Olark Chat | 21.878355 |
| Last Activity_Email Opened | 14.492187 |

Negative features

| | |
|---|---|
| Lead Origin_Landing Page Submission | -6.650706 |
| Last Notable Activity_Modified | -20.847089 |
| Specialization_Hospitality Management | -21.360161 |
| Do Not Email | -25.794517 |

# Determining Feature Importance

- Selecting Top 3 features which contribute most towards the probability of a lead getting converted

Last Notable Activity_Had a Phone Conversation
Lead Origin_Lead Add Form
Occupation_Working Professional

| | |
|---|---|
| Last Notable Activity_Had a Phone Conversation | 100.000000 |
| Lead Origin_Lead Add Form | 90.834614 |
| Occupation_Working Professional | 72.829897 |

# Conclusion

Conclusion After trying several models, we finally chose a model with the following characteristics:

- All variables have p-value < 0.05.
- All the features have very low VIF values, meaning, there is hardly any multicollinearity among the features. This is also evident from the heat map.
- The overall accuracy of test data set is 80.84% at a probability threshold of 0.34 on the test dataset is also very acceptable.

# Conclusion (contd….)

- Based on our model, some features are identified which contribute most to a Lead getting converted successfully.

- The conversion probability of a lead increases with increase in values of the following features in descending order:
- Last Notable Activity_Had a Phone Conversation
- Lead Origin_Lead Add Form
- Occupation_Working Professional
- Last Notable Activity_Unreachable
- Last Activity_Less Popular Activity
- Lead Source_Welingak Website
- Last Activity_SMS Sent
- Time on Website

# Conclusion (contd….)

- The conversion probability of a lead increases with decrease in values of the following features in descending order:

- Lead Origin_Landing Page Submission
- Last Notable Activity_Modified
- Specialization_Hospitality Management
- Do Not Email

# Conclusion (contd….)

- Another point to note here is that, depending on the business requirement, we can increase or decrease the probability threshold value with in turn will decrease or increase the Sensitivity and increase or decrease the Specificity of the model.

- High Sensitivity will ensure that almost all leads who are likely to Convert are correctly predicted where as high Specificity will ensure that leads that are on the brink of the probability of getting Converted or not are not selected.

*THANK YOU!!!!!*