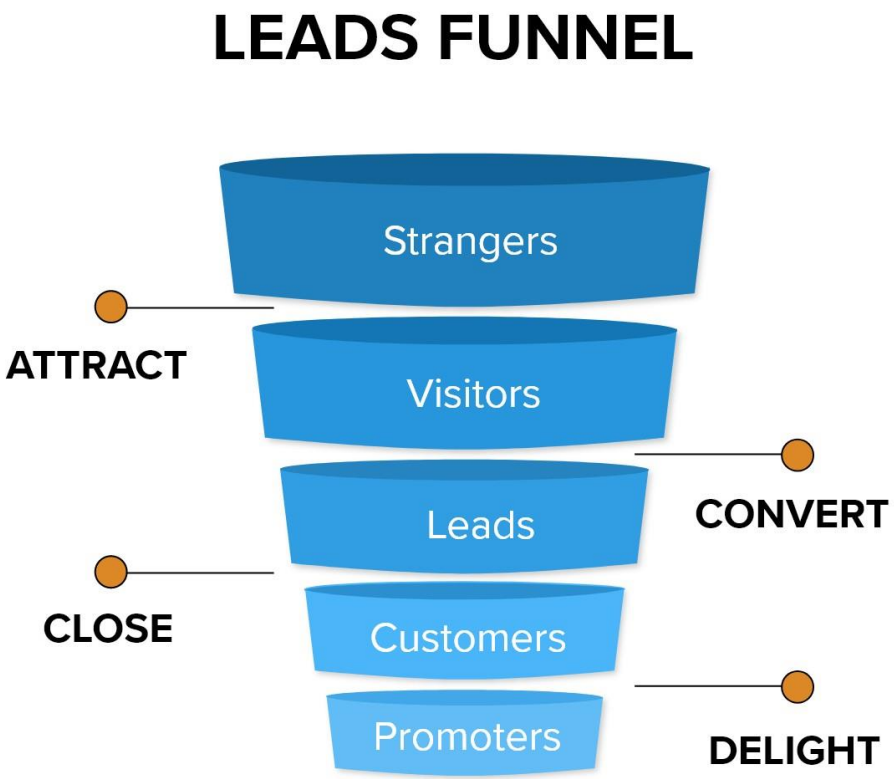
LEAD SCORING CASE STUDY

BY :

- ADRASHA SAHOO
- JOSHUA BAHIRVANI

PROBLEM STATEMENT



- An education company named X Education sells online courses to industry professionals.
- 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.
- The objective is to build a model to identify the hot leads and achieve lead conversion rate to 80%.

DATA INFORMATION

Information regarding the data:

- **Dataset used : Leads.csv**
- **Total number of customers present : 9240**
- **Total number of features : 37**
- **ML Model used : Logistic Regression**

- After initial analysis, we see that there are multiple factors that influence conversion rate
- We need to reduce the features to maximize the conversion rate.
- Current Conversion Rate =38.66%

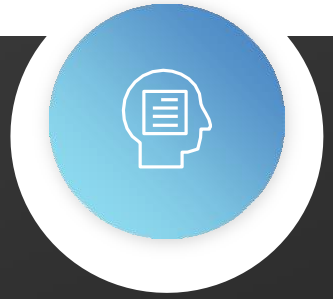
DATA CLEANING

Final list of features after cleaning

- Prospect ID
- Lead Origin
- Lead Source
- Do Not Email
- Converted
- Free copy of mastering the Interview
- Total Visit
- Total Time Spent
- Page View per Visit
- Last Activity
- Country
- Specialization
- Current Occupation
- What matters to choosing course
- City
- Last Notable Activity

Handling Select variable

- "Select" variable indicates that the user has not selected any option.
- We impute the same with null values.

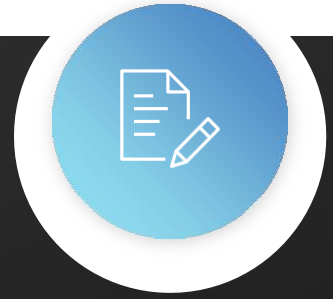


Dropping the Unique value Columns

- Dropped below unique value columns: 'Magazine', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque'

Treating Categorical data

- High Data Imbalance – Columns having high data imbalance must be removed.
- For e.g. : Category A has 98% , and Category B has 2% - This data is irrelevant to our analysis as one category is overpowering the other.
- Use best matrix such as mean, median or mode to impute the less missing values .

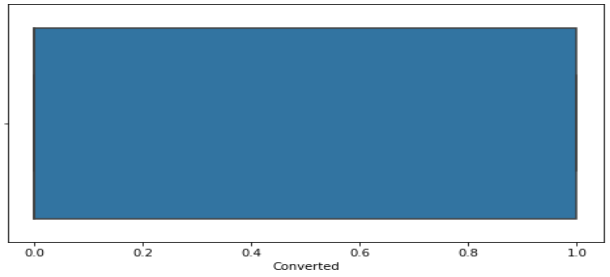


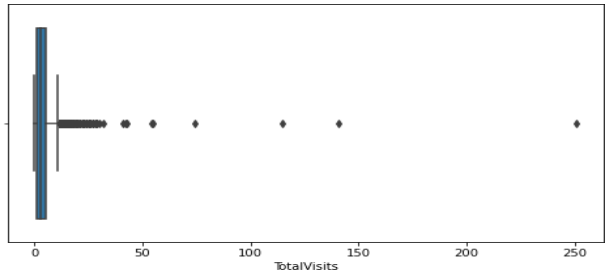
Dropping column with high null values.

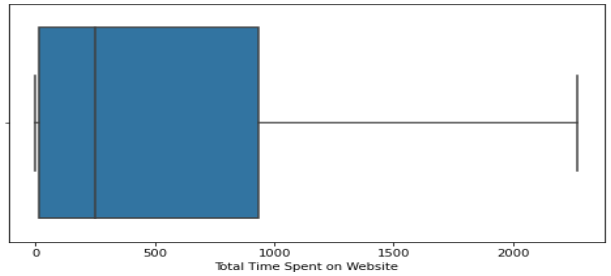
- Columns having null values greater than 45% does not have meaning to the data, hence we drop these columns
- For Specialization, we consider the column where people have not selected any value into one more column known as Not Specified and we use this for model building.

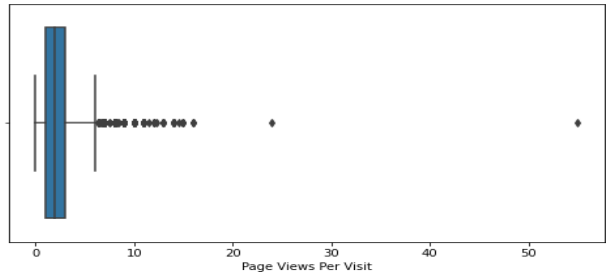
Outlier Treatment

Numerical variables – Checked the presence of outliers





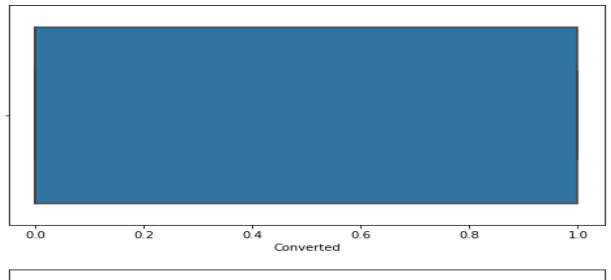


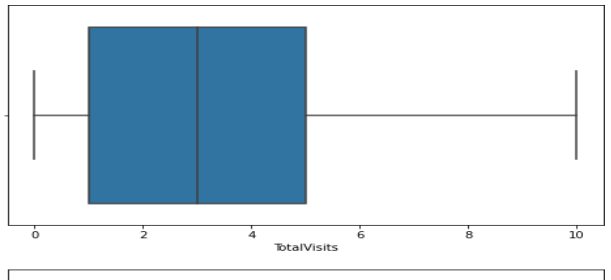


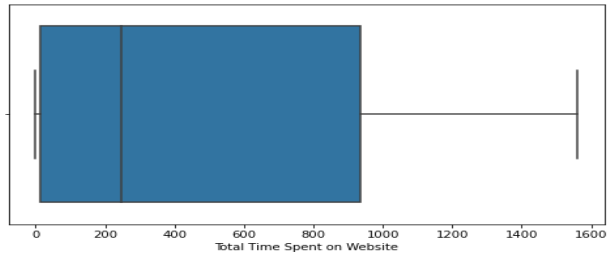
Inferences / Observation :-

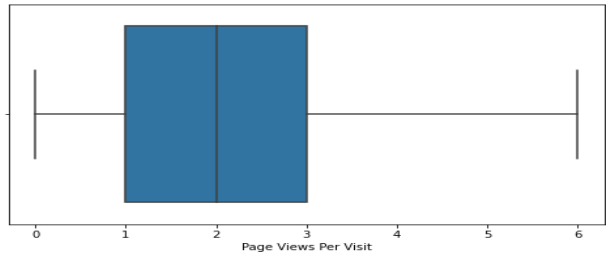
- There are so many outliers present in only two variables. Hence , the we have capped the outliers within the soft range.

Numerical variables –Outliers Capping









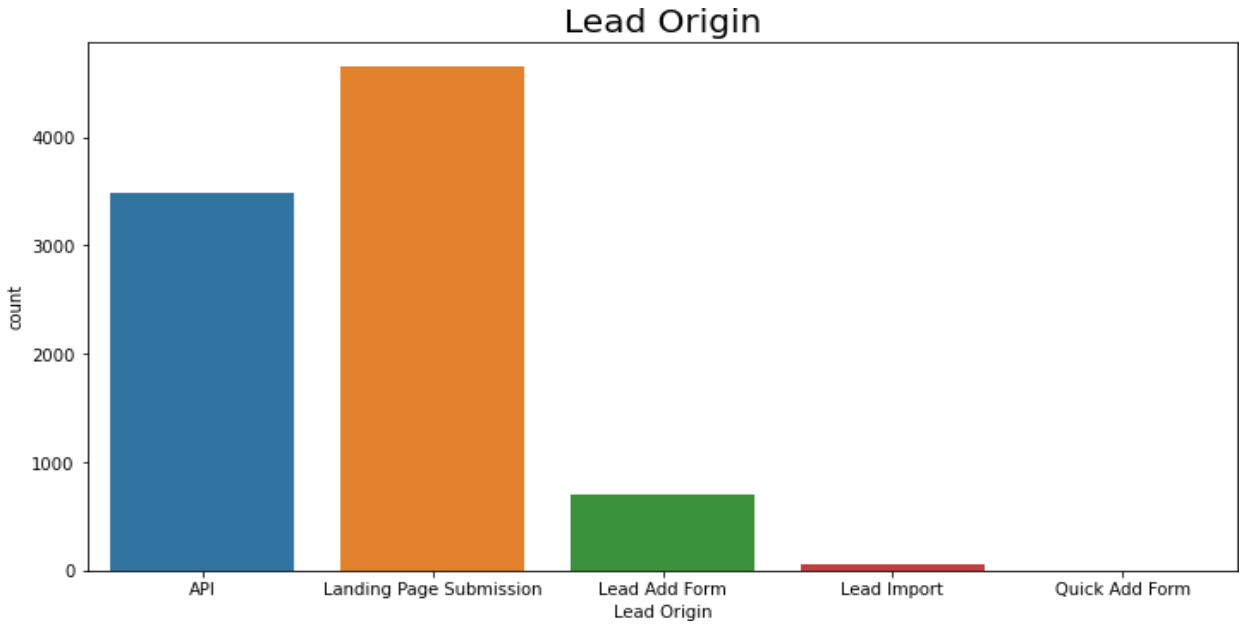
Inferences / Observation :-

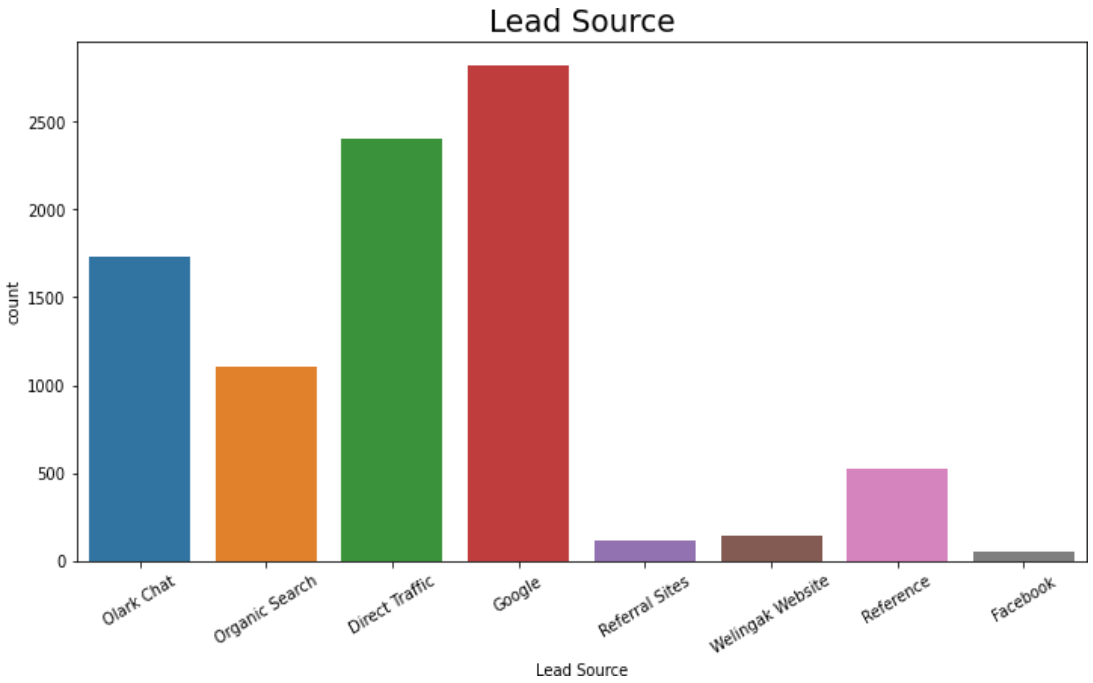
- Now there are no outliers as we have performed capping now we can proceed with the further cleaning.

Exploratory Data Analysis

Univariate analysis for categorical variables.

- Univariate analysis are carried out on 6 categorical variables i.e. Lead Origin, Lead Source, Last Activity, Specialization, What is your current occupation and City.

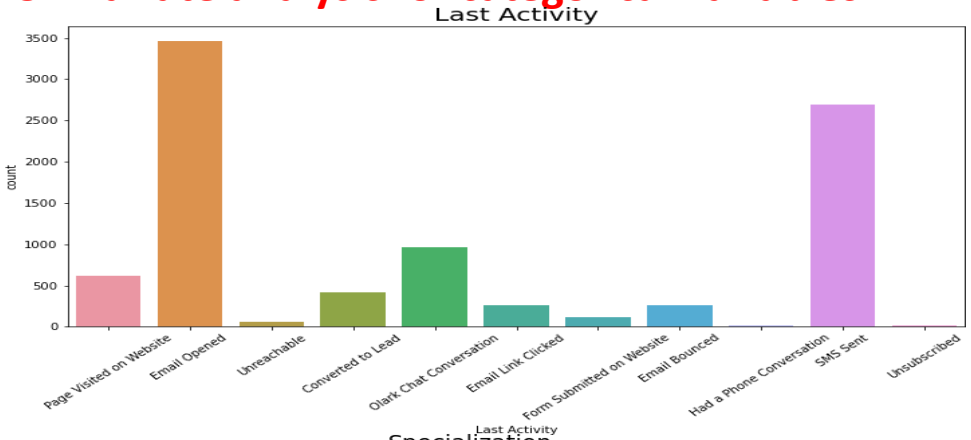




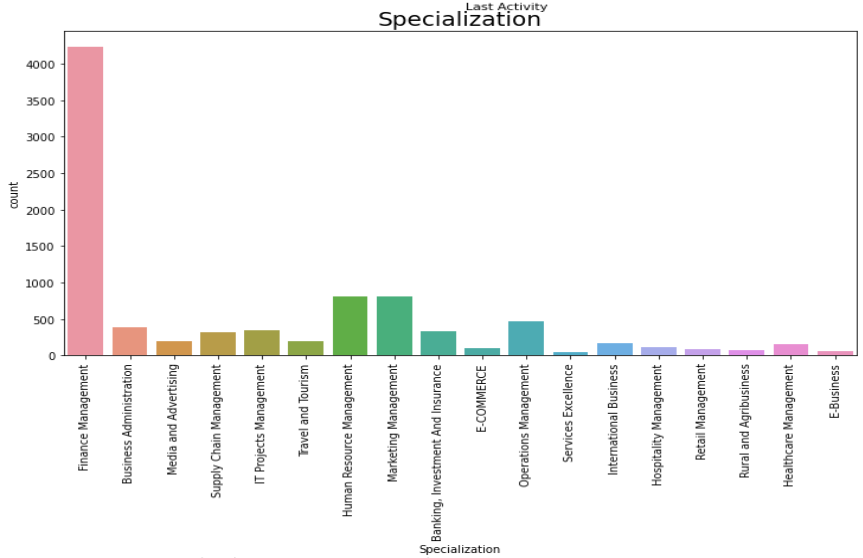
- Most of the customer is identified by Landing Page Submission.

- Most of the customer are from Google and Direct traffic to X education website.

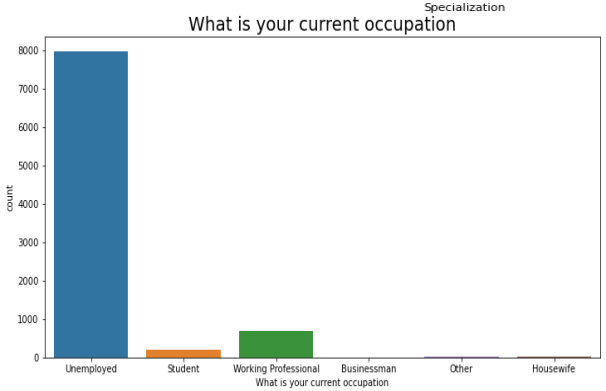
Univariate analysis for categorical variables.

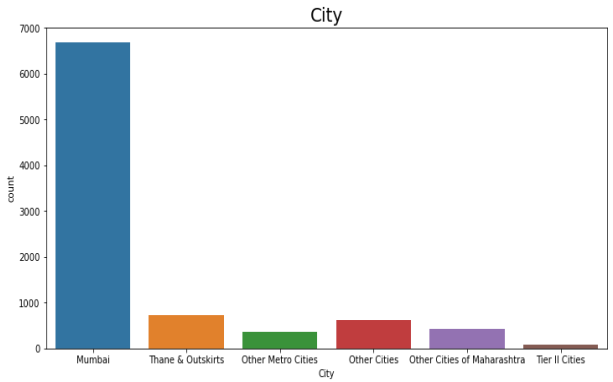


- Most of the customers have performed last activity as Email Opened and SMS Sent.



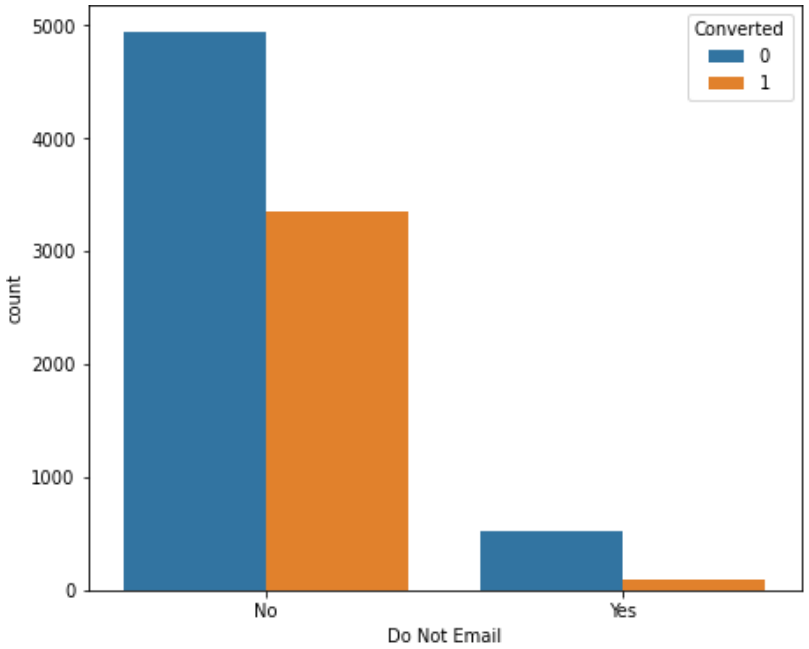
- Most of the customers are from Finance Management specialization and worked there before.

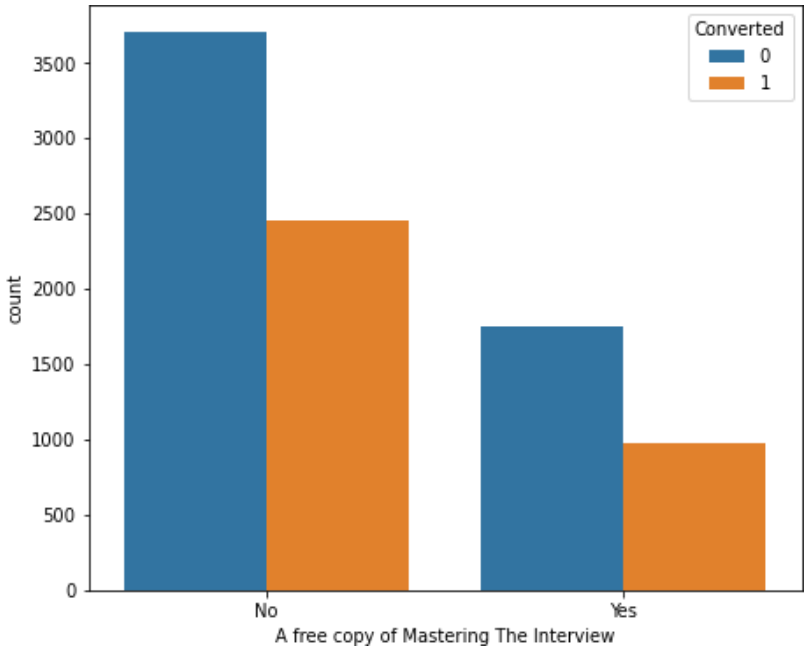




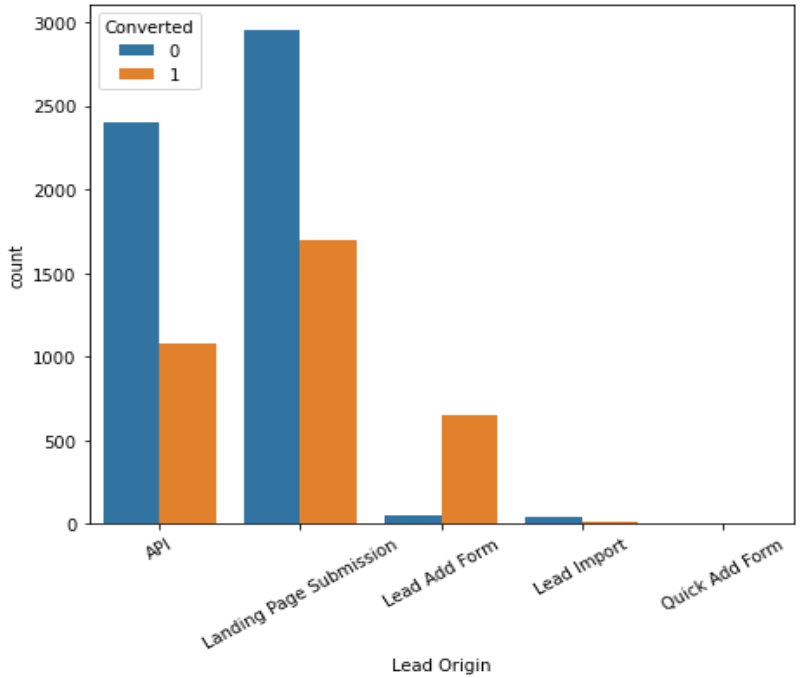
- Most of the customers are Unemployed.
- Most of the customers are from Mumbai City.

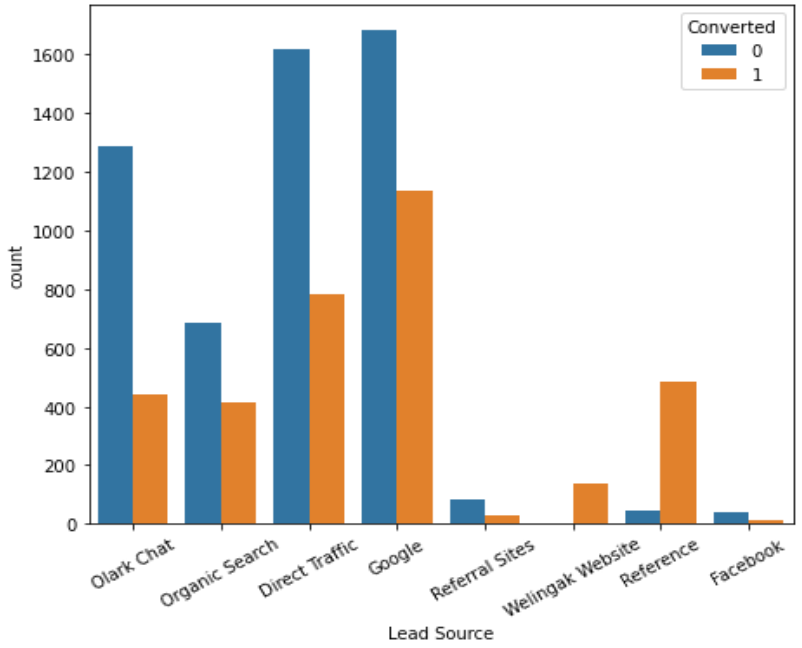
Univariate analysis for categorical variables with Target variable





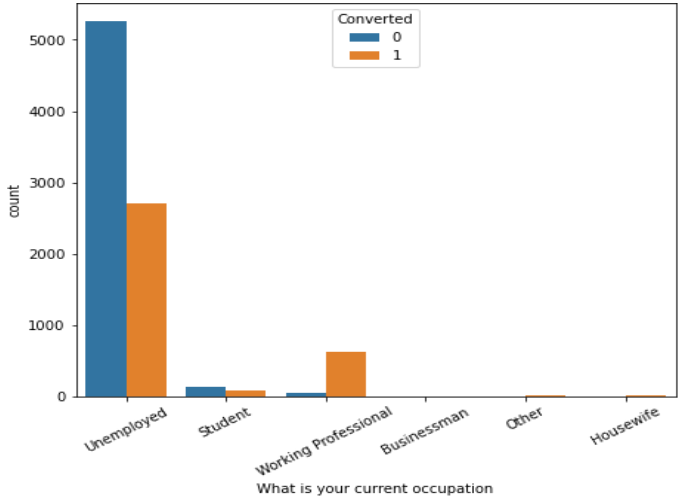
- Most of the leads do not want to be emailed about the course and also do notwant the free copy of mastering the interview.
- Those leads who do not want to be emailed have high chances of getting converted.

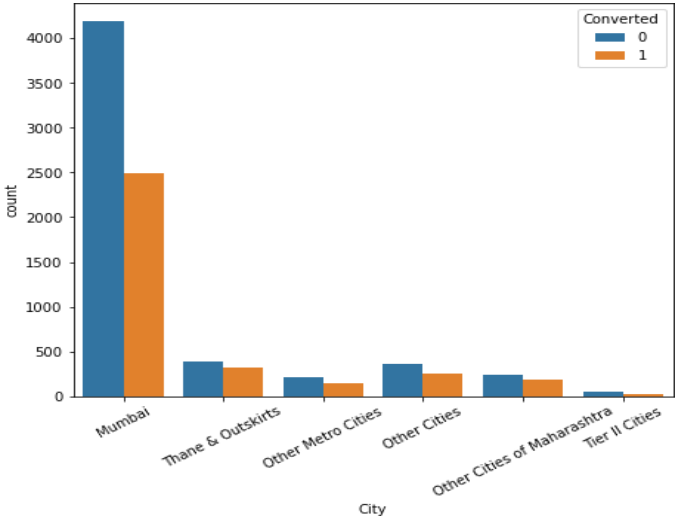




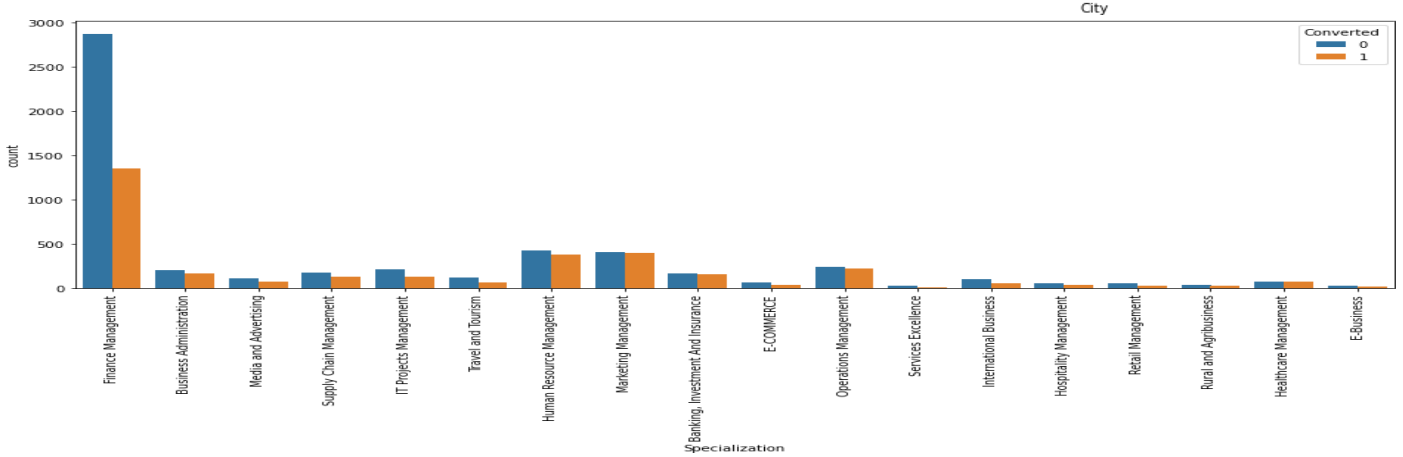
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- API and Landing Page Submission bring higher
- Google, Direct Traffic and Olark Chat bring higher number of leads as well as conversion.

Univariate analysis for categorical variables with Target variable

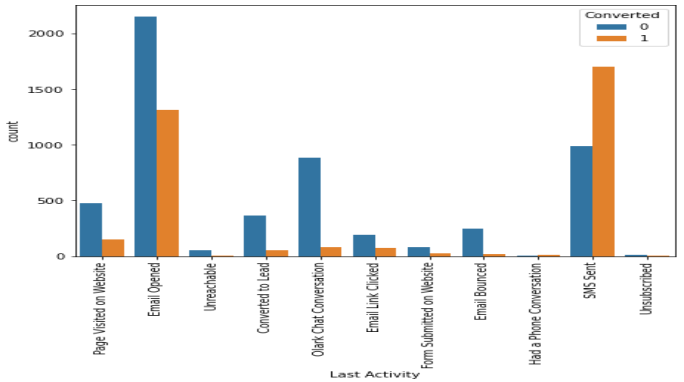


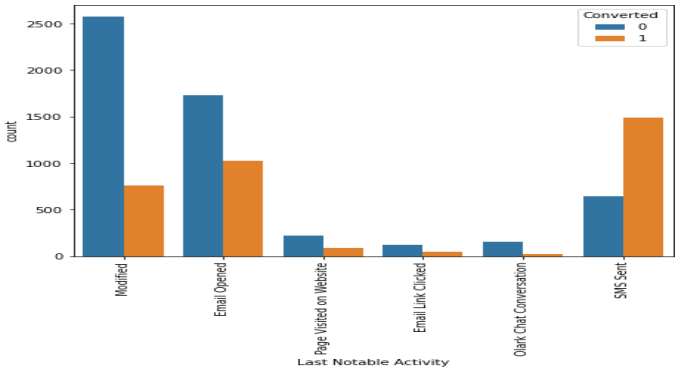


- Higher number of leads as well as conversion from Unemployed category.
- Mumbai has the highest number of leads as well as conversion.



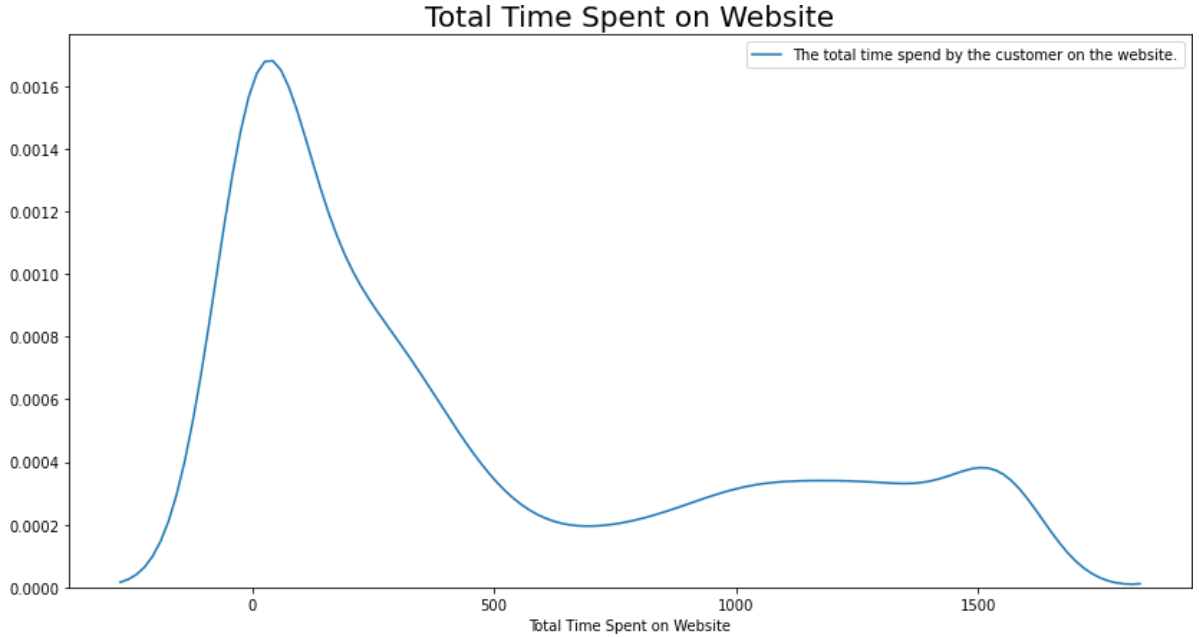
- Finance Management, Human Resource Management, Marketing Management, Operations Management are showing reasonably good results in terms of count of leads as well as conversion.



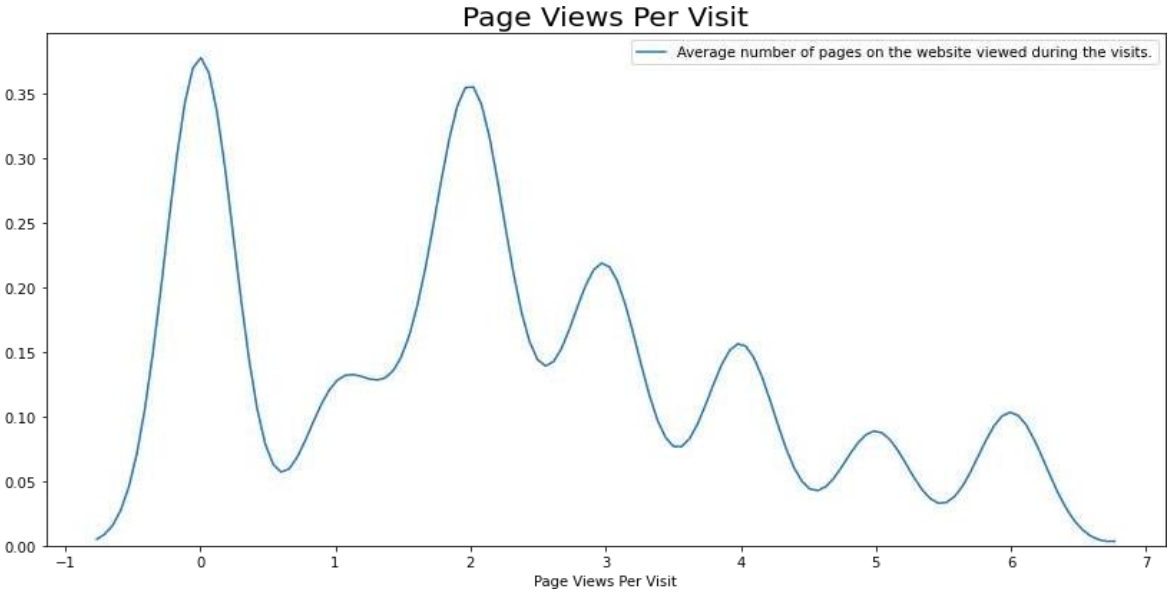


- Although the count is high for 'Email Opened', but the highest conversion rate from 'SMS Sent' Category.
- High Conversion rate is for 'Email Opened' and 'SMS Sent' Category.
- Lead count is highest for 'Modified' and 'Email Opened' category.

Univariate analysis for Numerical variables

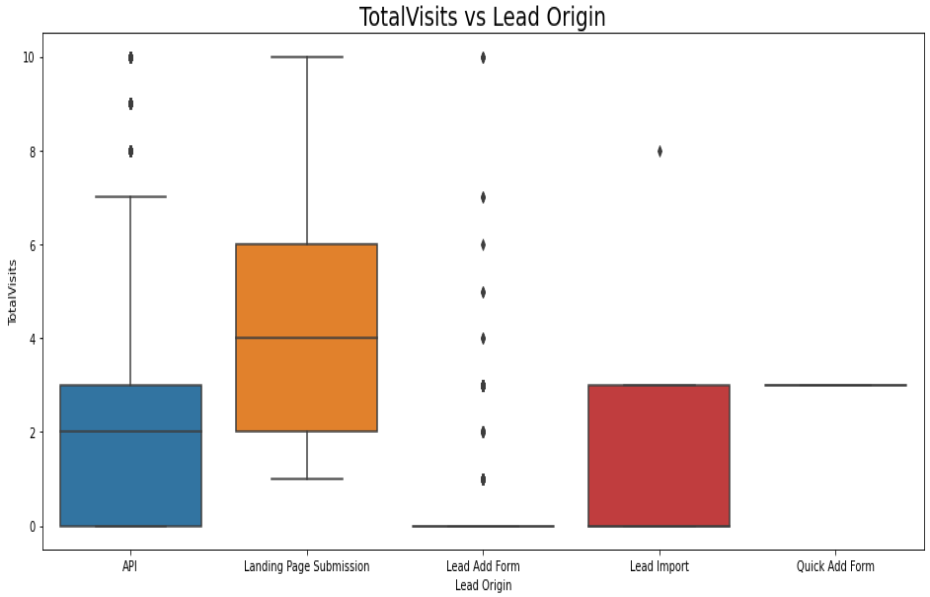


- Maximum number of customers are having a time spent between 0 to 500 seconds.
- There are very low customers who spends more time on website.

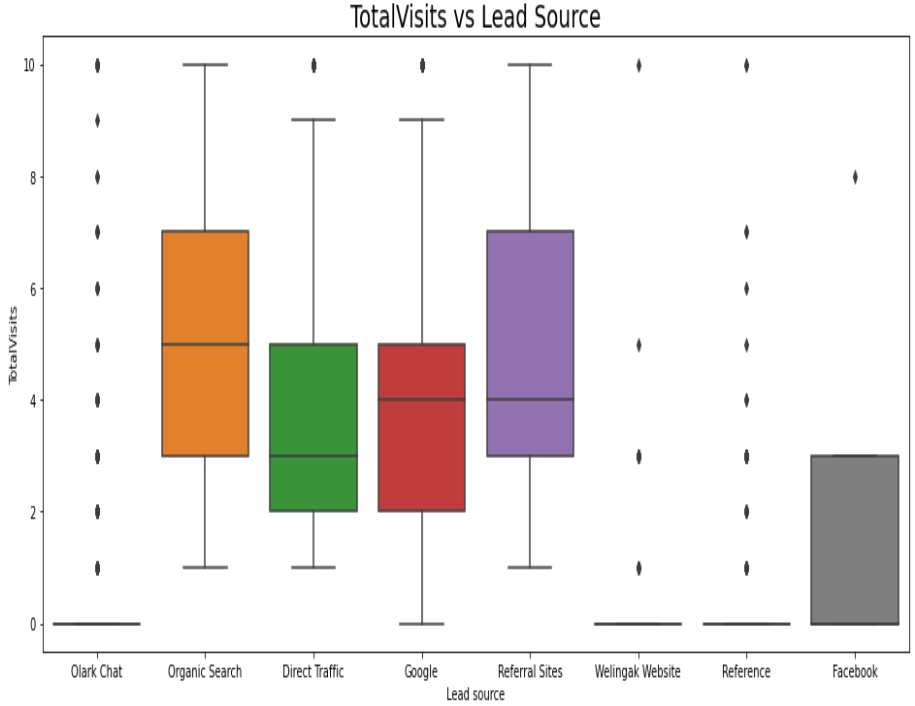


- Most customers are visited 4 pages and there are very less customers who visited more than 4 pages.
- Maximum number of page views is 2 to 3.

Bivariante Analysais for Continuos - Categorical variables

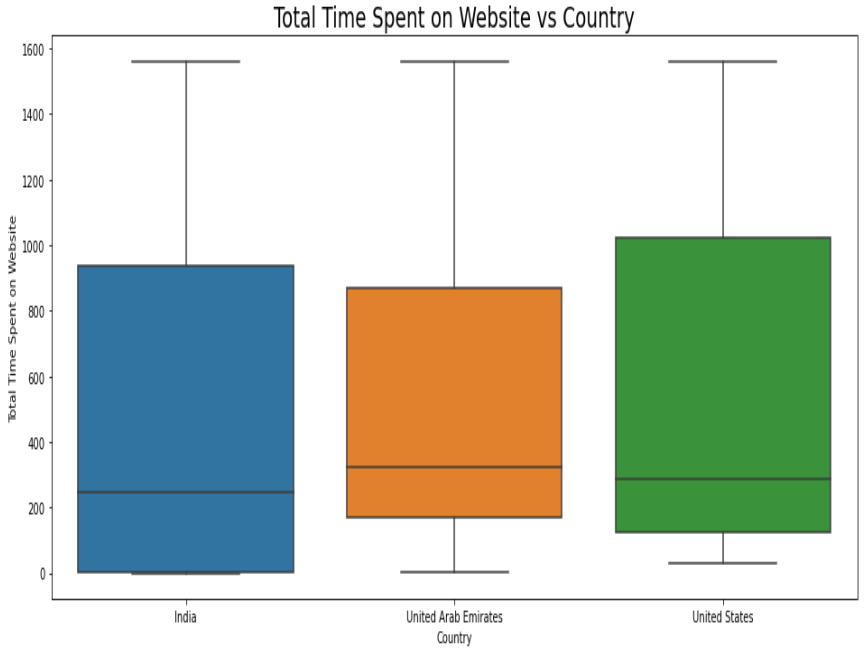


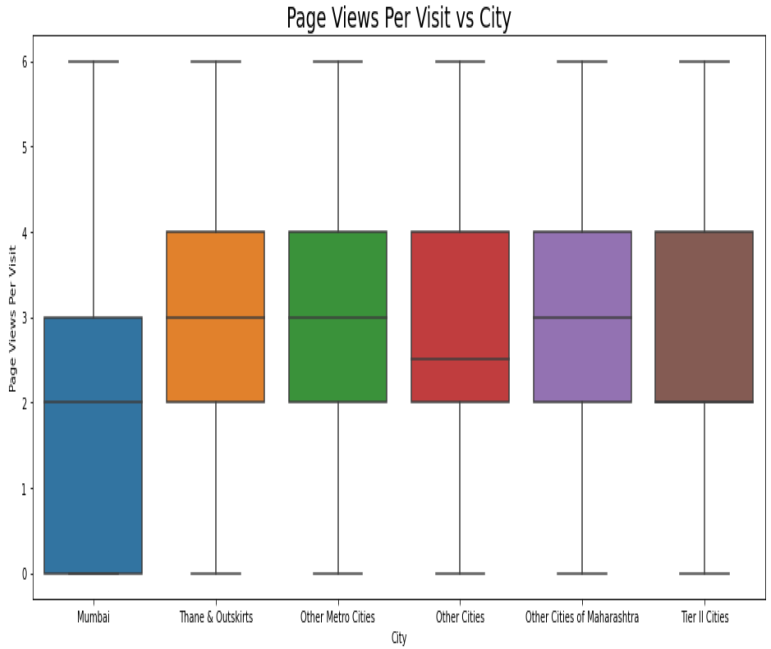
- The customers who have origin as Landing Page Submission have higher amount of total visits followed by API and Lead import.

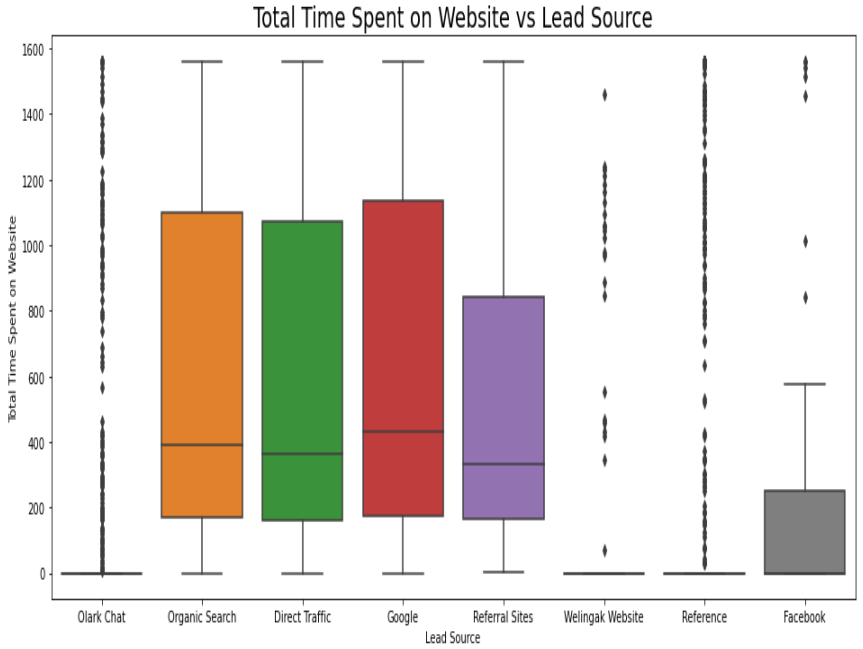


- The customers who have source as Organic Search have higher amount of total visits followed by Referral Sites Lead Source, Direct Traffic and Google.

Bivariante Analysais for Continuos - Categorical variables





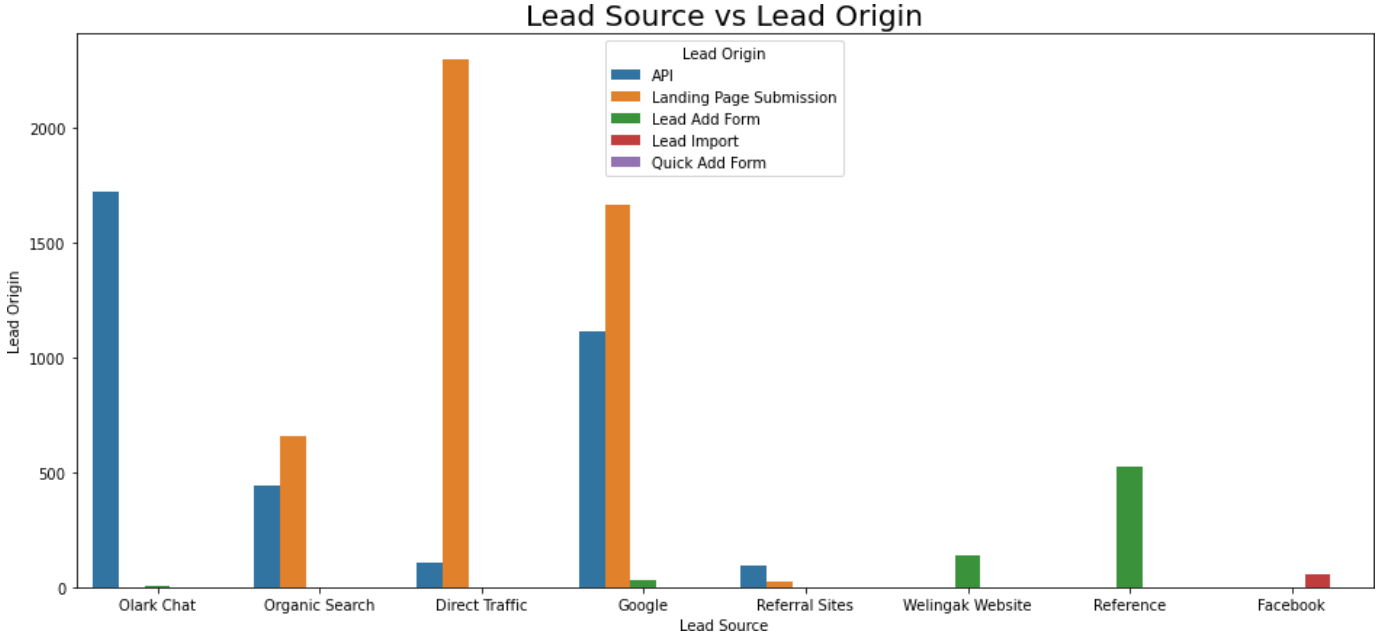


- Almost every country has same amount of time spent on website

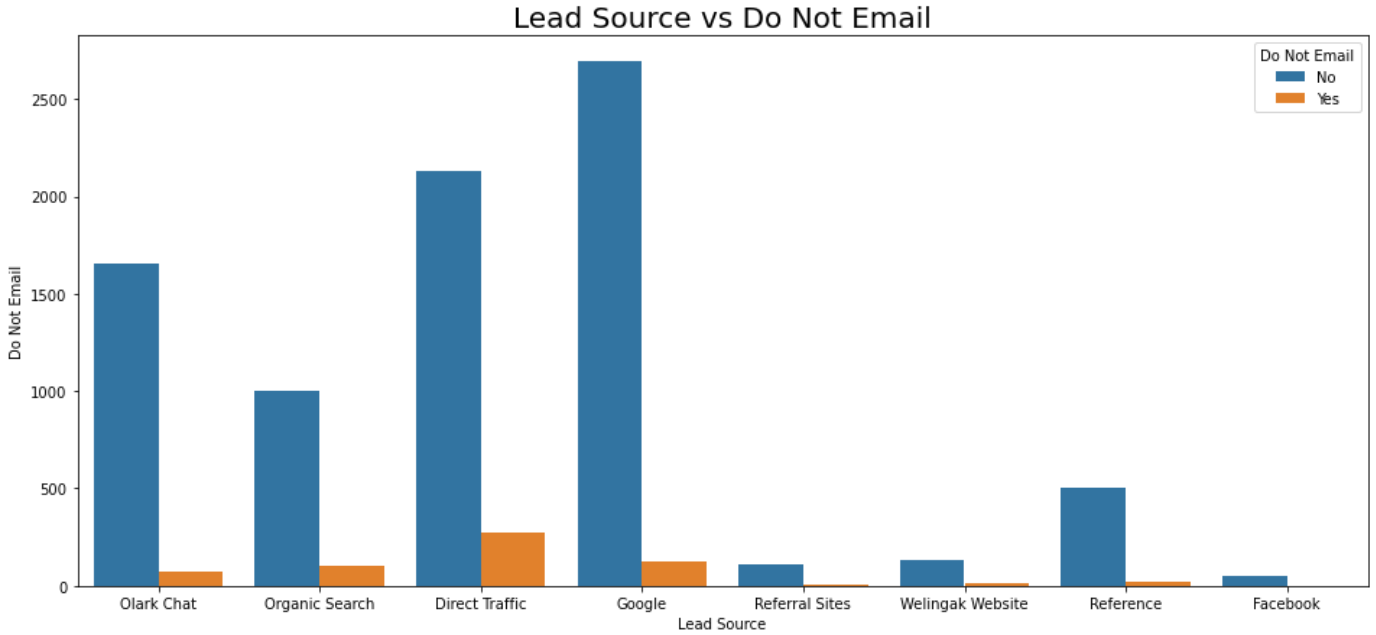
- Every city has almost same number of page views per visit except Mumbai has less number of Page Views Per Visit.

- The customer who has lead source as "Google", "Direct Traffic", "Organic search" has higher time spent of website.

Bivariante Analysais for Continuos - Categorical variables

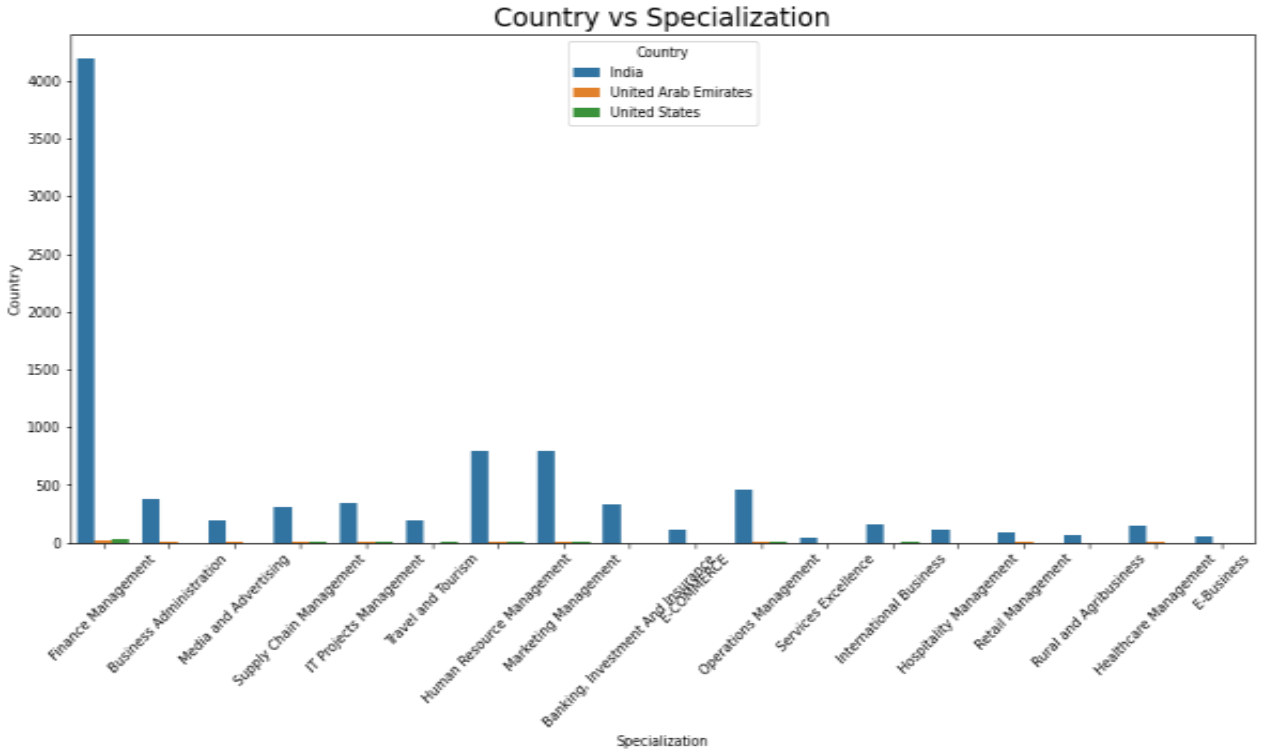


The Olark Chat source has API as its origin most of times.
The most customer which are from Direct Traffic source has origin as Landing Page Submission.
The most customer which are from Google source has origin as Landing Page Submission.

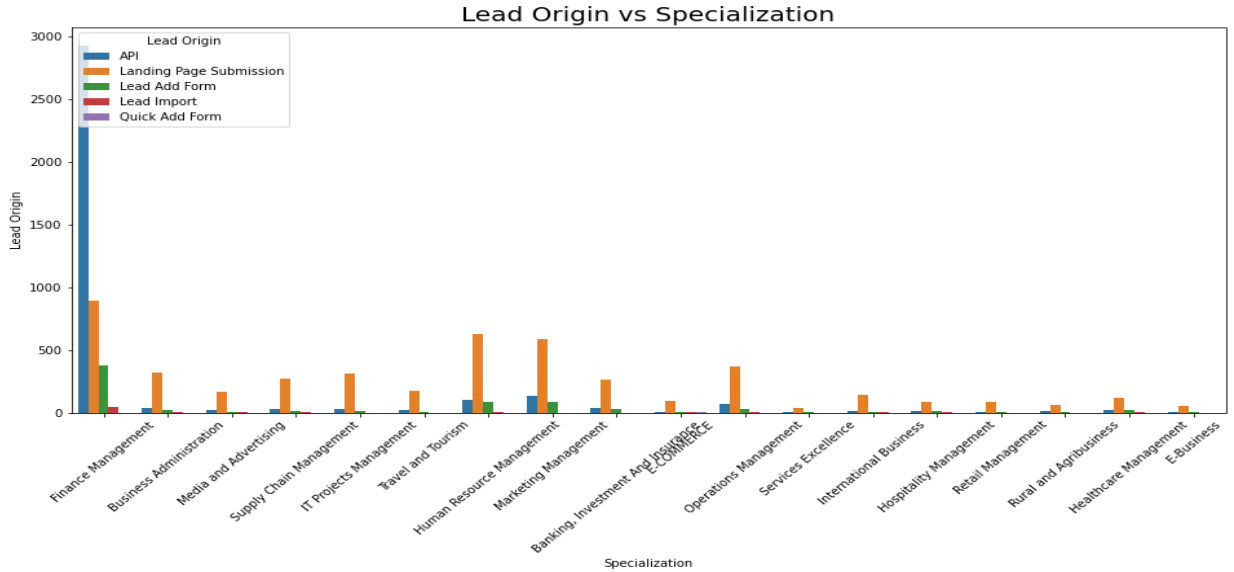


- The customer who has specialization as Finance Management has Origin as API and Landing Page Submission.
- From every specialization most of the customers has origin as Landing Page submission.

Bivariante Analysais for Continuos - Categorical variables

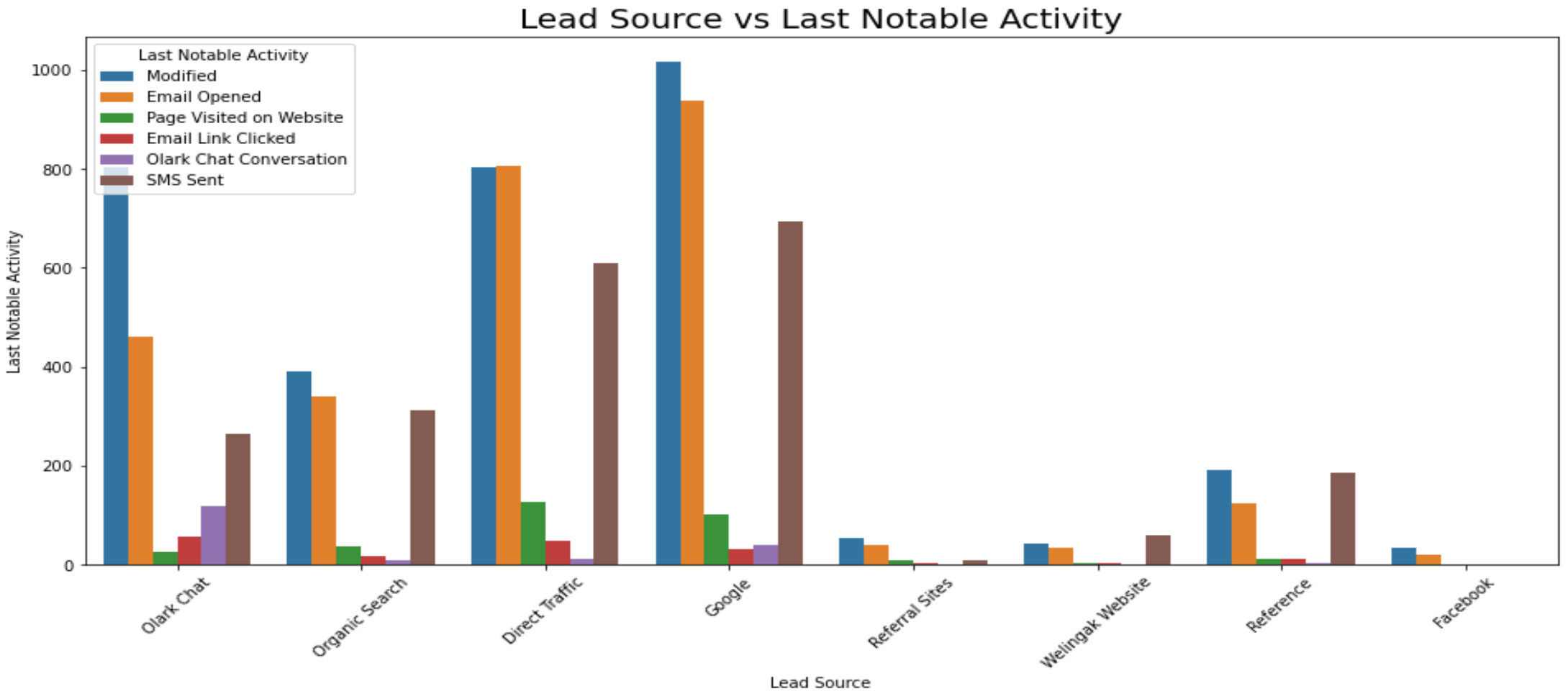


- Customers which are from India has highest Specialization as Finance Management.



- The customer who has specialization as Finance Management has Origin as API and Landing Page Submission.
- From every specialization most of the customers has origin as Landing Page submission

Bivariate Analysis for Categorical- Categorical variables.



- Most of the customers who have Source as Google, Direct traffic, Olark Chat and Organic Search have Last notable activity as Modified and Email Opened.

DATA PREPROCESSING BEFORE MODEL BUILDING

There are a few basic steps that needs to be followed for preparing the data before model building.
Model building:

- *Libraries used: StandardScaler()*
- *RFE is used to perform variable selection effectively and to eliminate the insignificant columns*

1

Droping 'Last Activity' Columns:

In our data frame we have one column which is sales team column. This column is generated once the sales team get into call with the student so we need to drop it.

2

Binary Mapping:

- In this step we converted the binary column into zero's and one's. All the categorical columns are replaced by their respective dummies

3

Dummy Variable Creation:

- We need to create dummy variables for all the categorical columns as they enable us to use a regression equation on multiple groups.

4

Test Train Split:

- Division of data into test data and train data to check the stability of the model.
- We have randomly sampled 70%of the data as the test data and 30%of the data as test data.
- Random State =100

5

Scaling:

- Division of Train Data into X and Y where X has all the features and Y has the target variable – Converted.
- We perform scaling to normalize the data within a particular range
- Technique : Standard Scaler

MODEL BUILDING STRATEGY

Use RFE for Feature Selection.

Running RFE with 15 variables as output.

The statsmodel library is used to build the logistic regression model

Building Model by removing the variable whose p- value is greater than 0.05and VIF value is greater than 5.

Model –6 is our final model :

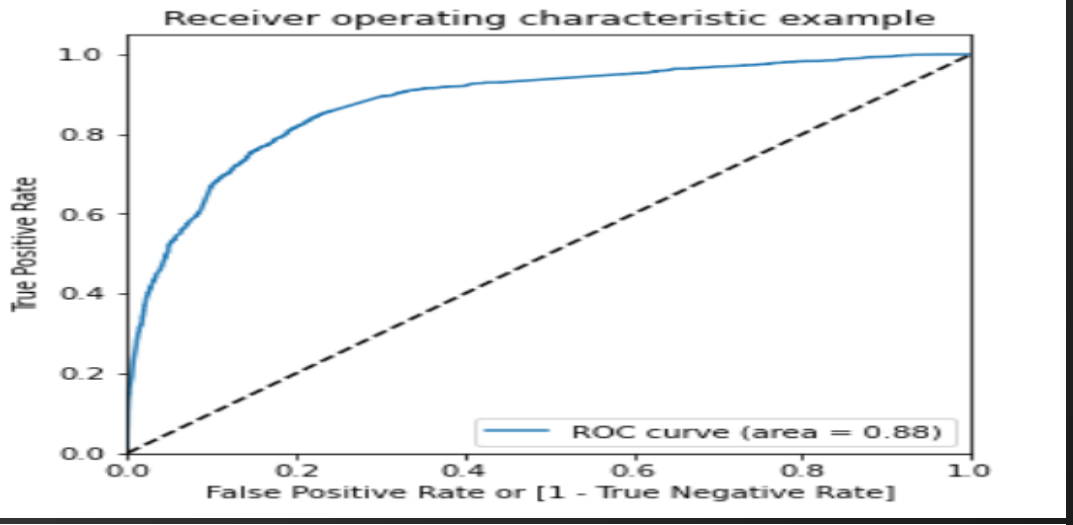
- All p-values <5%- Hence they are highly significant
- All VIF values are < 5%- Hence the dependency of variable with another is tolerable.
- Final model has 10 features in tota

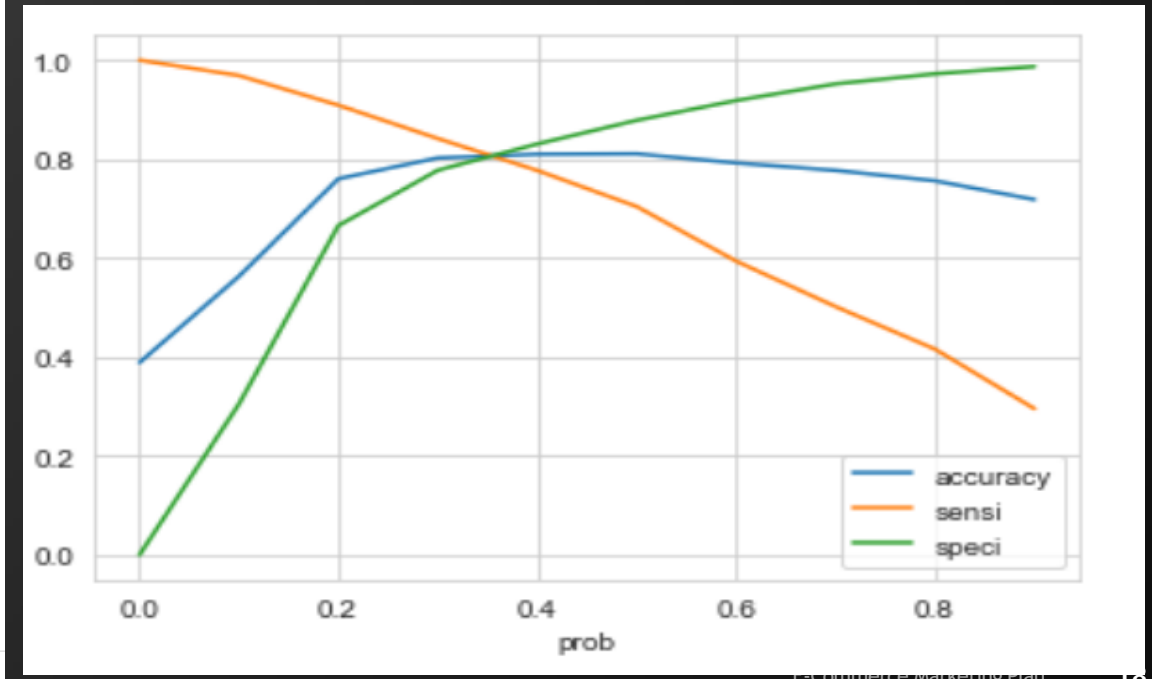
ROC CURVE AND PROBABILITY CUT OFF

AUC =>0.88 whereas Optimal Cutoff Probability =>0.36

ROC Curve and Optimal Cutoff Point

- ROC Curve represents how much the model is able to distinguish between the classes.
- AUC – Area under the curve represents that it is distinguishing the 1's and 0's correctly.
- On plotting the ROC curve for our data we see that, AUC is around 0.88 which means at around 88% of the times, the model is able to distinguish the 1's as 1's and 0's as 0's.
- AUC of 0.88 is found to be very stable model.
- When we plot the sensitivity, accuracy and specificity of the model together, the optimal cut off point is found to be at 0.35. This means that at 36 %probability, the sensitivity and specificity are found to be balanced.
- With probability =0.36, we predict y-values with X-Train, in such a way that, any conversion prob >36%is said to be converted to a lead.





Model Performance Parameters – Train Set v/s Test Set

Train Set

- **Accuracy : 80.76%**
- **Sensitivity : 80 . 43%**
- **Specificity : 80 . 97%**

Test Set

- **Accuracy : 80.34%**
- **Sensitivity: 81 . 04%**
- **Specificity: 79.91 %**

- **The Accuracy and sensitivity value after model building process is found to be greater than 80% as required.**

Summary

1.The customer/leads who fills the form are the potential leads.
2.We must majorly focus on working professionals
3.It's always good to focus on customers, who have spent significant time on our website.

Based on analysis, defining the results and conclusion.

The features which are most mattered in lead conversion are : (Arranging from most important to less important by comparing the coefficient.)

1. **Lead Origin_Lead Add Form**
2. **Lead Source_Welingak Website**
3. **What is your current occupation_Working Professional**
4. **Last Notable Activity_SMS Sent**
5. **Do Not Email**
6. **Total Time Spent on Website**
7. **Last Notable Activity_Olark Chat Conversation**
8. **Lead Source_Olark Chat**
9. **Last Notable Activity_Email Opened**
10. **Lead Origin_Landing Page Submission**

OTHER BUSINESS RECOMMENDATIONS

Based on our model , X Education Company should focus on the following from a business perspective:

1. Increase online engagement and better the Search engine optimization (SEO) for Google and social media while reducing print media engagement

2. Work to increase Total Visit on website by making website engaging and self explanatory via chat bots

3. Improve Olark Chat as this affected conversion rates negatively

THANK YOU!