Lead Scoring Case Study

Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
X Education gets a lot of leads, its lead conversion rate is very poor.
X Education want the help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

Essentially, the company wants —
•Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
•Use of model for future needs.

By Team:
Sachin Kumar
Sachin Shinde
Sachin Jain

Solution Approach:

- **Data handling and cleaning**
  - Reading and understanding the data
  - Handling missing values
    - Delete columns with missing values >40%
    - Impute missing values as per requirement
  - Checking and handling Outliers

- **Performing EDA,**
  - For categorical variables, analyze the count/percentage plots.
  - For numerical variable, describe the variable and analyze the box plots

- **Creating Dummy Variables**
  - For categorical variables with multiple levels, create dummy features (one-hot encoded)

- **Test-Train Split**
  - The next step is to split the dataset into training an testing sets.

- **Feature Scaling**
  - Now there are a few numeric variables present in the dataset which have different scales. So let's go ahead and scale these variables.

- **Running Model and Feature Selection Using RFE**

- **Model Evaluation**

- **Calculate accuracy sensitivity and specificity for various probability cutoffs**

- **Making Predictions on the Test Set**

Data Cleaning details:

List of columns drop due > 40% missing values
- “How did you hear about X Education”
- “Lead Quality”
- “Lead Profile”
- “Asymmetrique Activity Index”
- “Asymmetrique Profile Index”
- “Asymmetrique Activity Score”
- “Asymmetrique Profile Score”

Impute missing values:
- “Specialization” - “Missing_Spec”
- "What is your current occupation“ - "Missing_Occup“
- "What matters most to you in choosing a course“ - "Missing_matter“
- “Tags” – “Missing_Tags
- “City – “ Missing_City

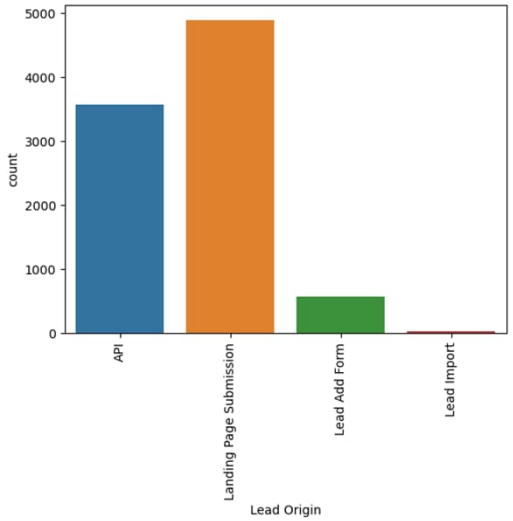
Removing missing rows from below columns:
- “Lead Source”
- “TotalVisits”
- “Page Views Per Visit”
- “Last Activity”

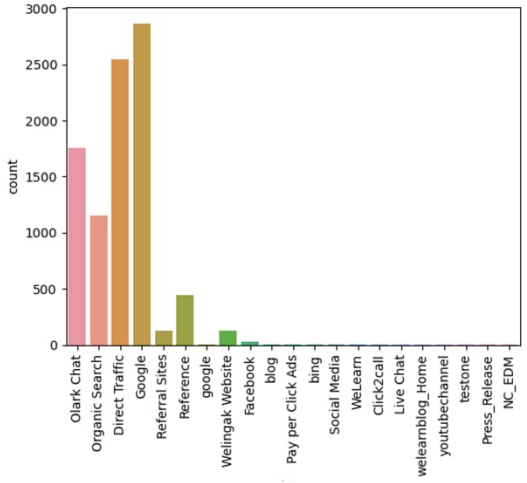
Dropping some columns which seems to be of least importance and not essential for building a model
- "Update me on Supply Chain Content“
- "Get updates on DM Content“
- "I agree to pay the amount through cheque“
- "Prospect ID“
- "Do Not Email“
- "Do Not Call“
- "Lead Number“
- "City“
- "Country“
- "Tags"

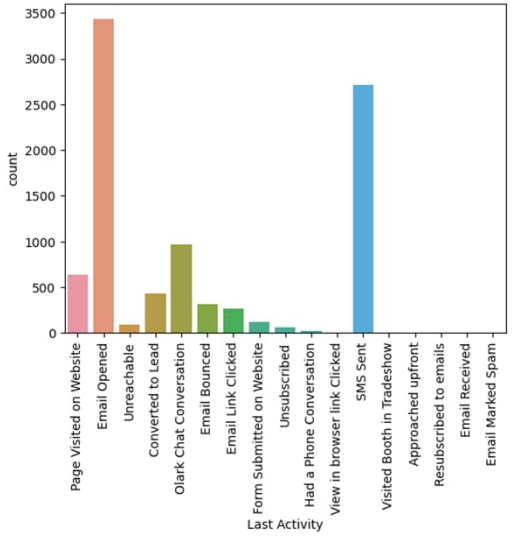
Data count after data cleaning:
Columns : 20
Rows : 9074

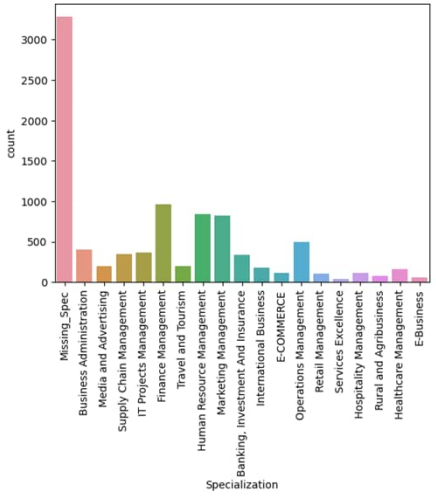
Performing EDA

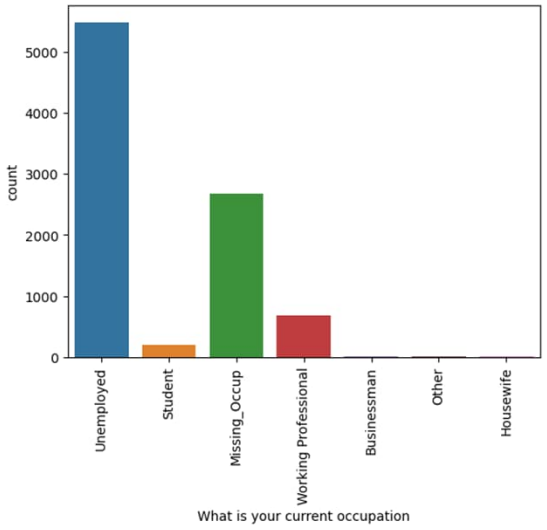
- For categorical variables, analyze the count/percentage plots.
- For numerical variable, describe the variable and analyze the box plots

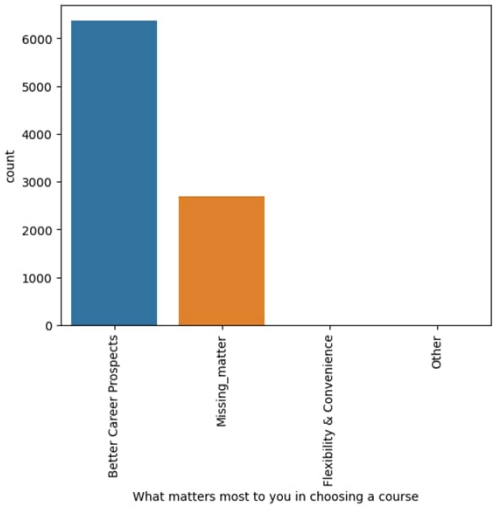


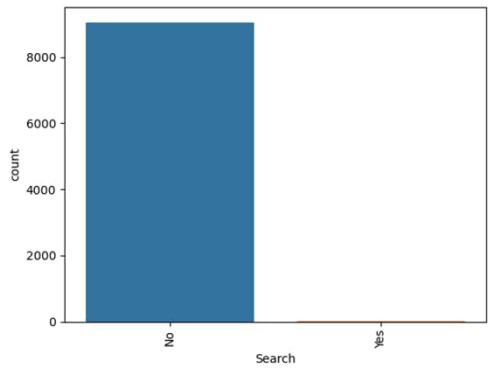


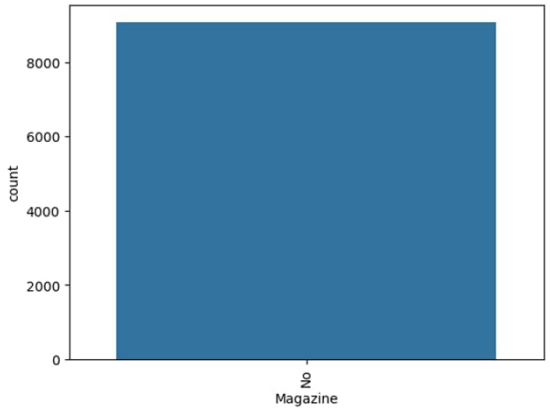


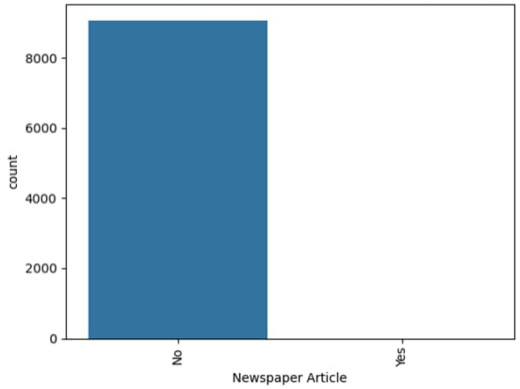


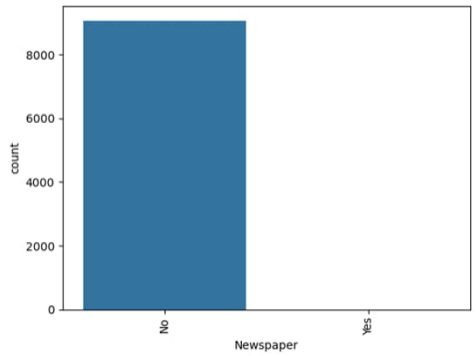


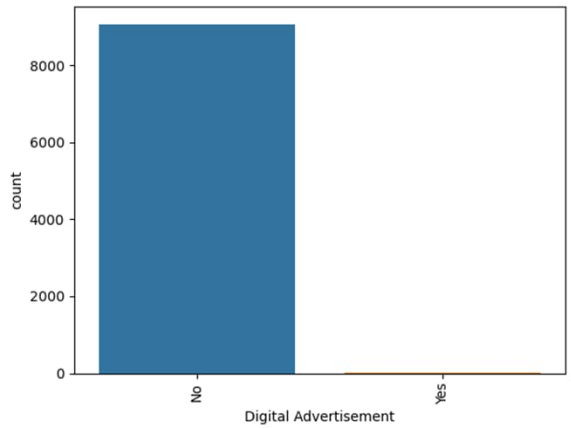


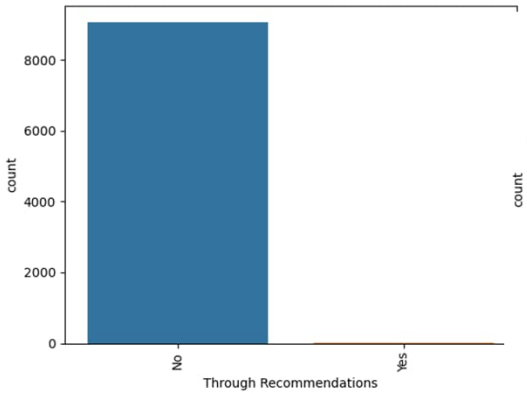

count
8000
6000
4000
2000
0
No
Yes
Search

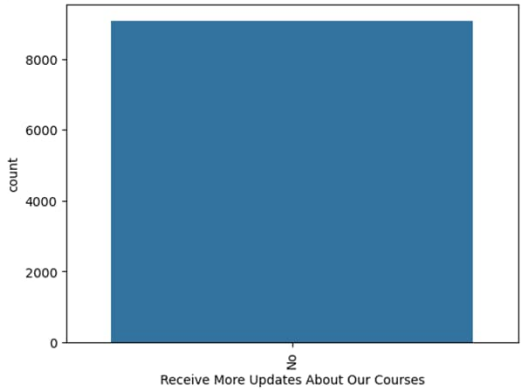

count
8000
6000
4000
2000
0
No
Magazine

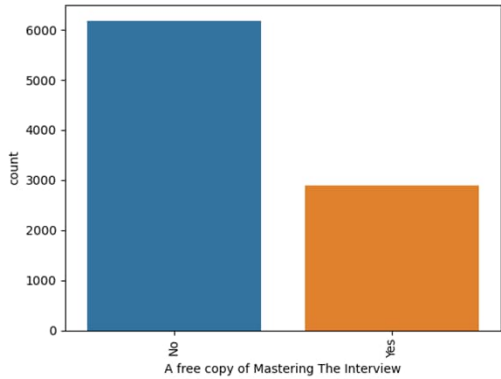

count
8000
6000
4000
2000
0
No
Yes
Newspaper Article

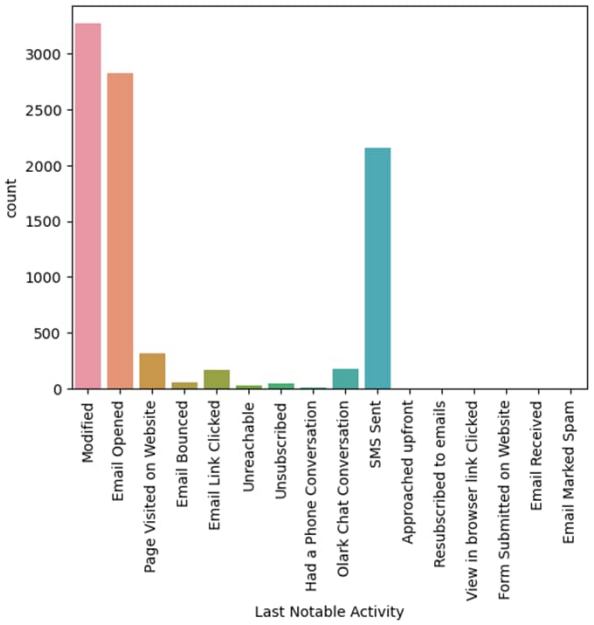

count
8000
6000
4000
2000
0
No
Yes
Newspaper

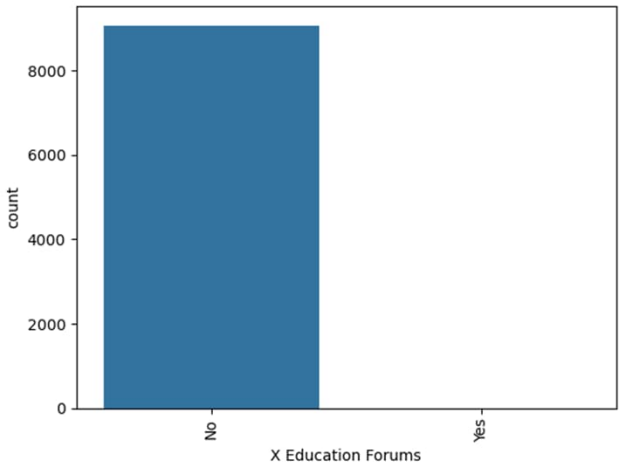

count
8000
6000
4000
2000
0
No
Yes
Digital Advertisement

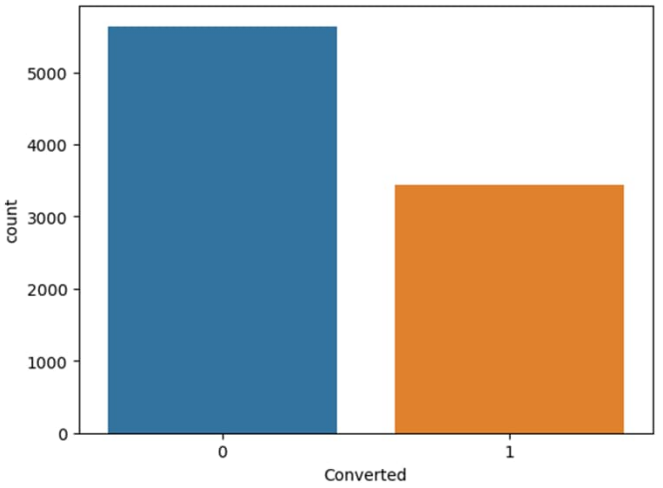

count
8000
6000
4000
2000
0
No
Yes
Through Recommendations


count
8000
6000
4000
2000
0
No
Receive More Updates About Our Courses

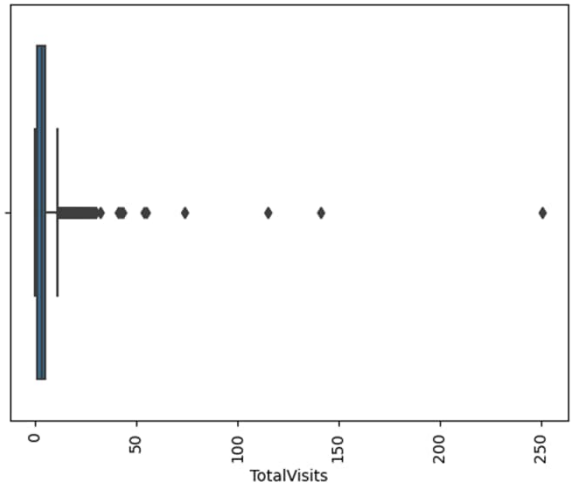

count
6000
5000
4000
3000
2000
1000
0
No
Yes
A free copy of Mastering The Interview

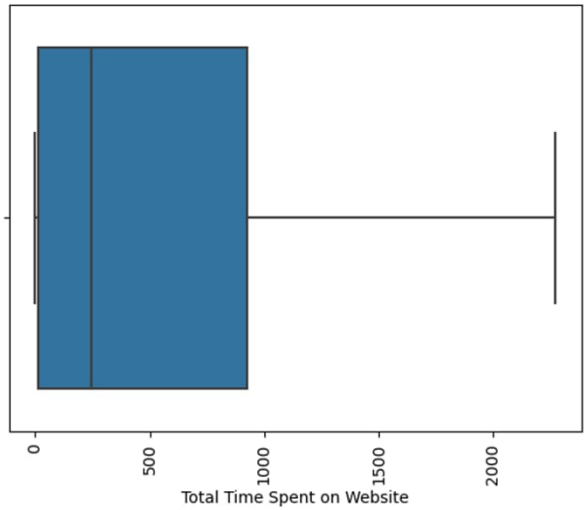


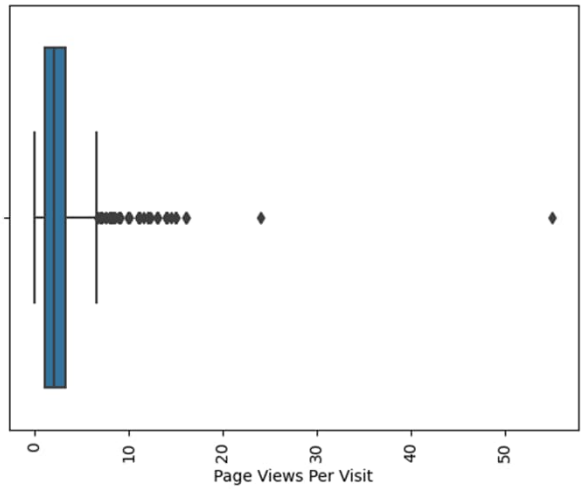


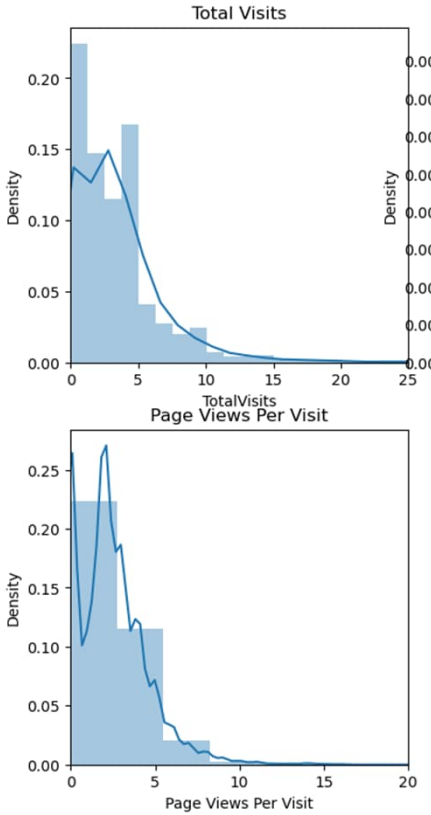


- **Conclusion:-**
  - •With this plot we came to conclusion that we can drop a few columns with one unique variable "Magazine", "Receive More Updates about Our Courses"
  - •More than 50% of the people have converted
  - •Lead Source is "Google"
  - •Most Promising last activity is either Opening an E-mail or is SMS
  - •lead Country is India , However there is a massive croud which has not specified anything.
  - •Finance . HR , Marketing and Operations are the few which have high scale , however a majority of people have not specified their specialization
  - •Unemployed People are likely to search more alojng with people for Better Career Prospects
  - •Mumbai is the city where major students enroll
  - •Mangzine,Newspaper Article,Newspaper,Digital Agvertisiment,Through Recommendations, Receive More Updates About Our Courses these can be droped



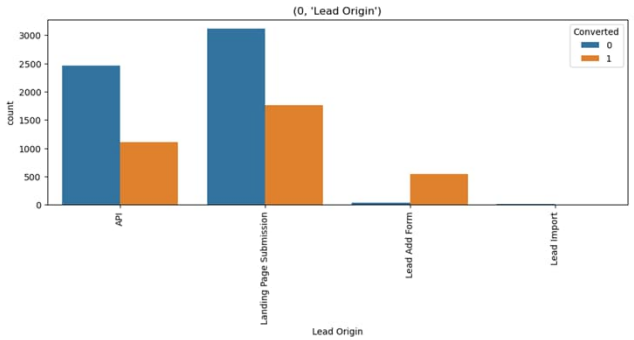


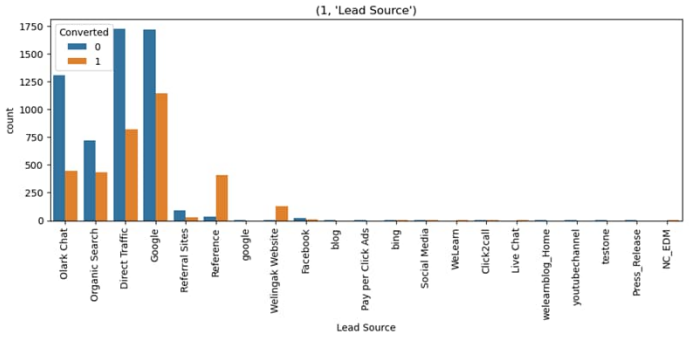


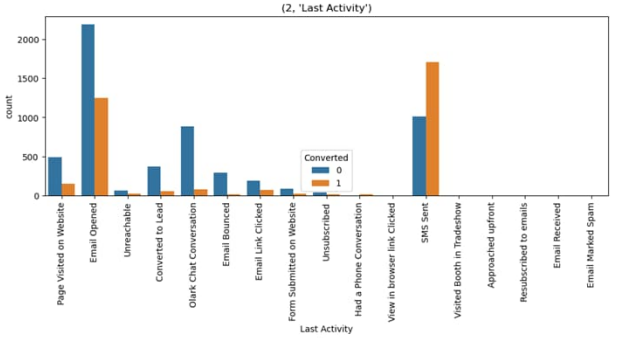


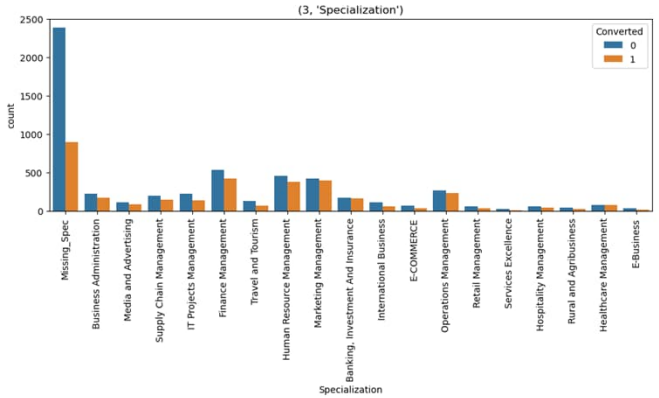
- **Handling Outliers**
  - Lets look at the respective percentile values and check if we can see if there is any sudden increase in the frequency

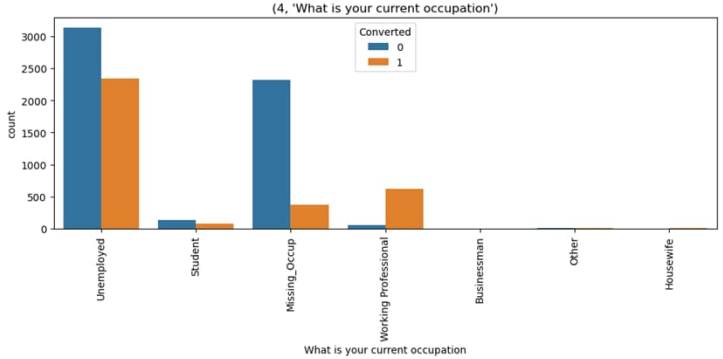
BI VARIANTE ANALYSIS

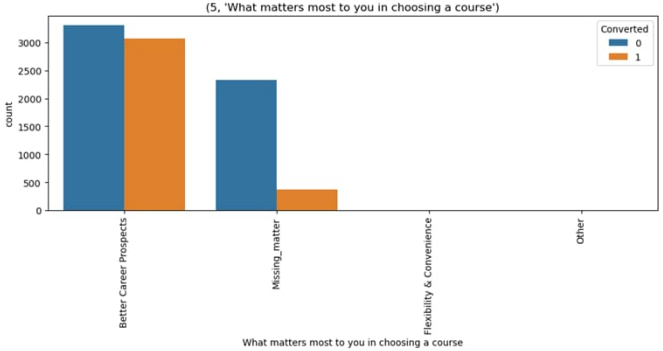


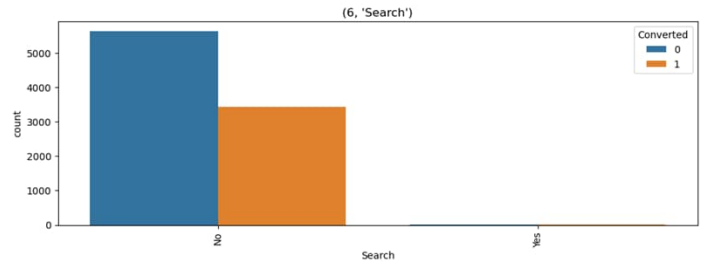


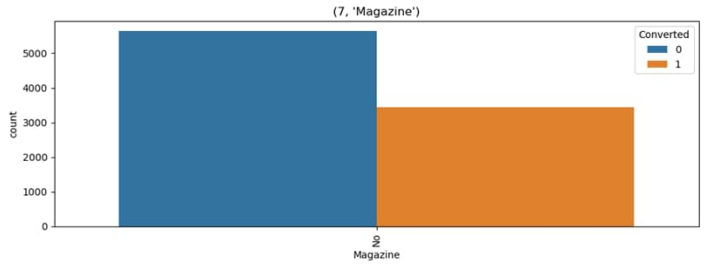


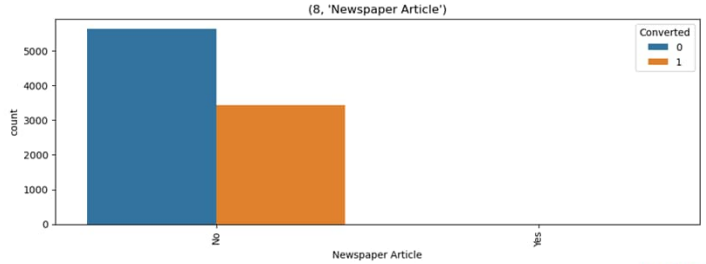


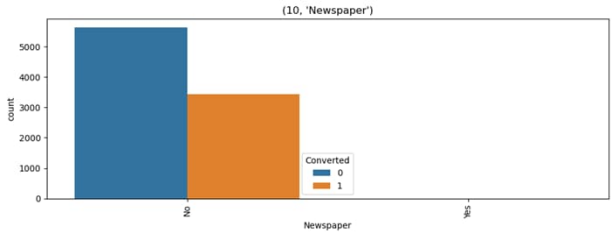


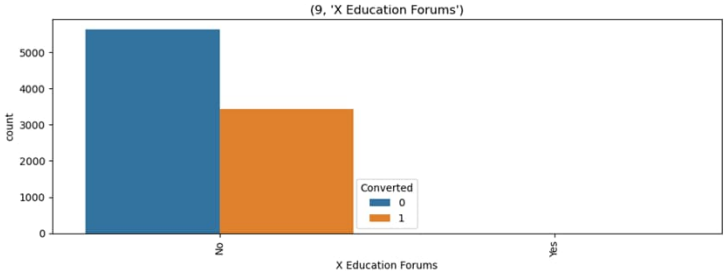


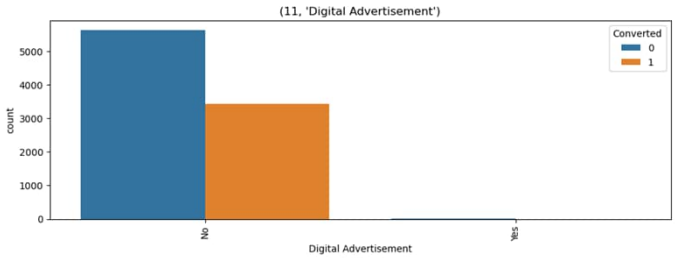


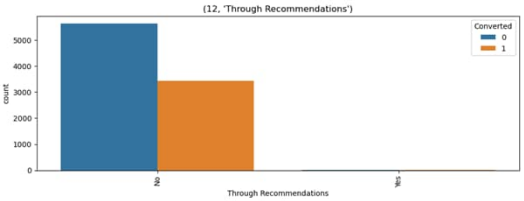


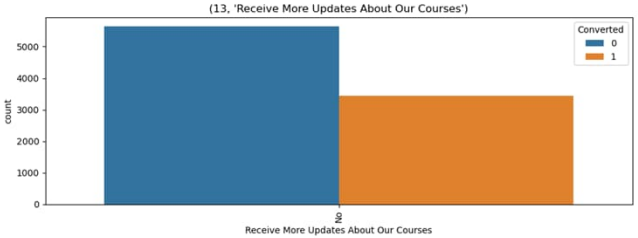


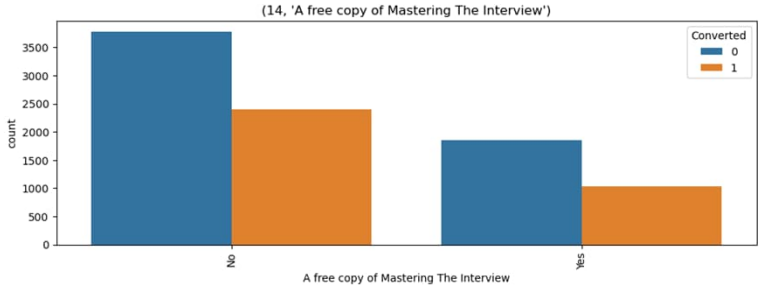


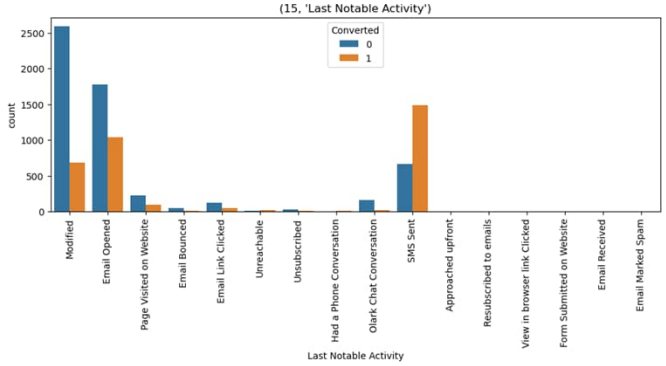








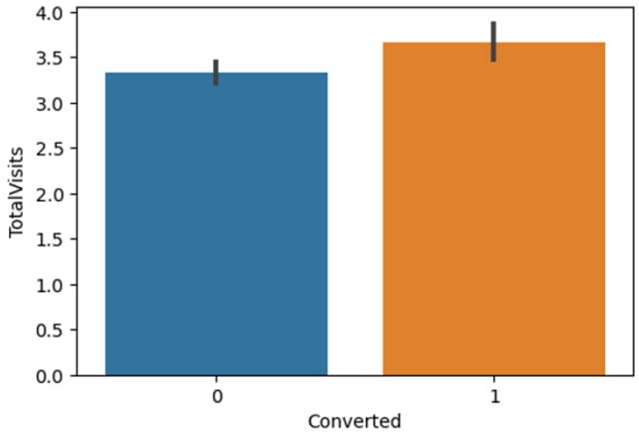


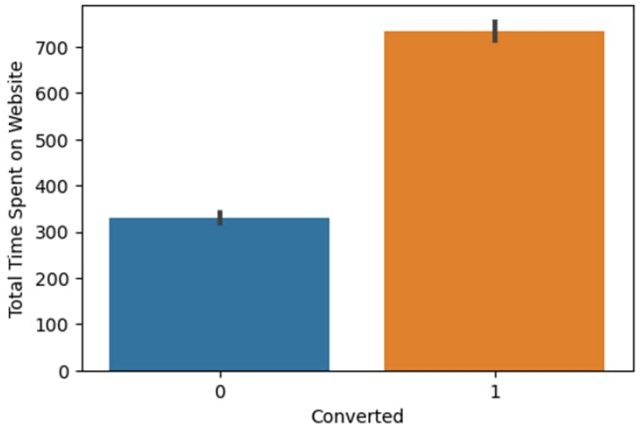


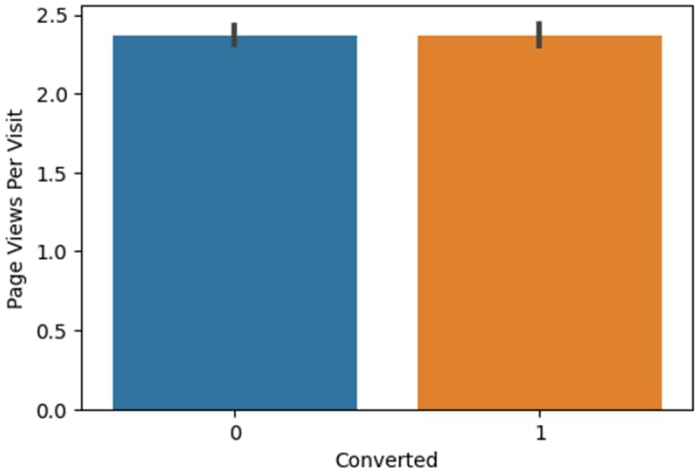
Conclusion

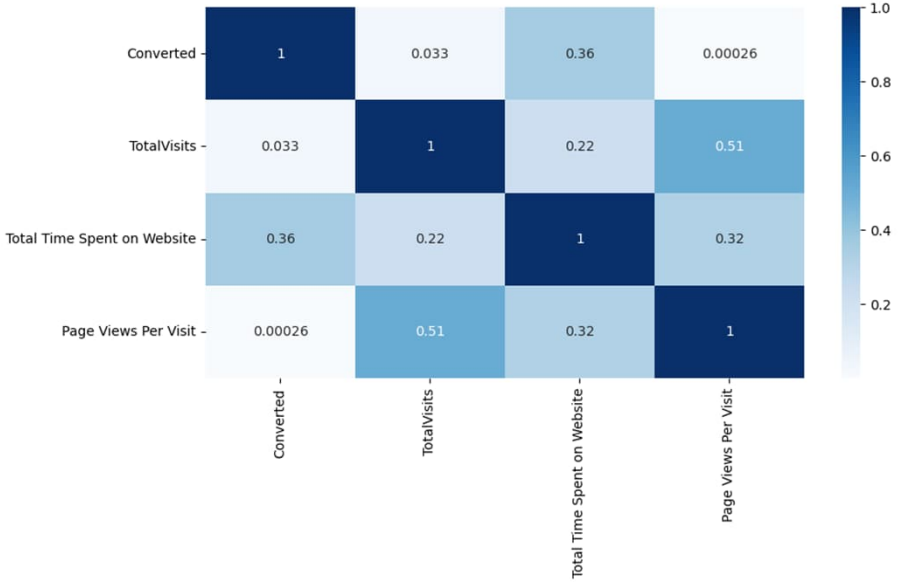
•From Lead Origin it can be seen that the maximum conversion happened from Landing Page Submission
•From Lead Source It can be seen that Google Search landed the maximum conversion
•From Last Activity SMS Sent had more conversion
•From Country India has high rate of conversion
•Finance Management has the high rate of conversion
•More conversion happend with people who are unemployed.
•Better Career Prospects is the leading cause of conversion
•Conversion rate is high on leads who are not through search
•Since "Newspaper Article" column now has only one value for all rows - "No" , it is safe to drop this column
•Since "X Education Forums" column now has only one value for all rows - "No" , it is safe to drop this column
•Since Newspaper column has only one row with "Yes" as the value and further since this lead did not get converted and rest of all the values are "No", we can safely drop the column
•Since "Magazine" column now has only one value for all rows - "No" , it is safe to drop this column
•Since "Newspaper Article","Digital Advertisement","Through Recommendations","Receive More Updates About Our Courses"
•Reverting after reading the email has the most conversion
•Conversion rate is high on leads who do not want a free copy of Mastering Interviews
•From Last Activity SMS Sent had more conversion

Checking For Numerical Variables:









Conclusion

•The conversion rated were high for Total Visits, Total Time Spent on Website and Page Views Per Visit

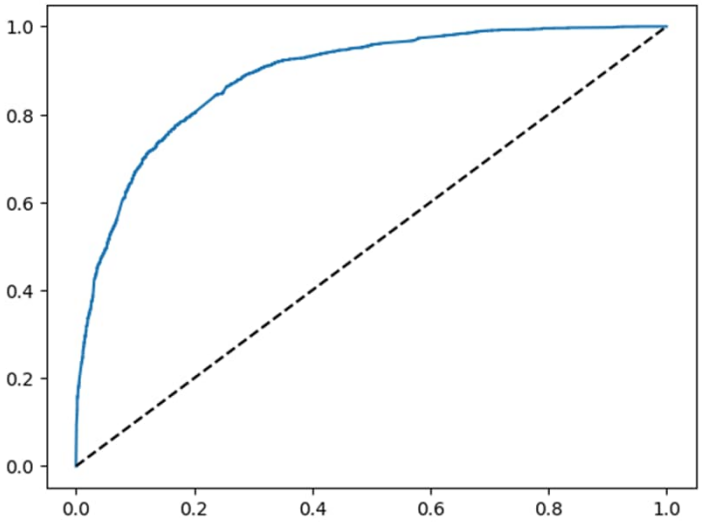
Model Building:

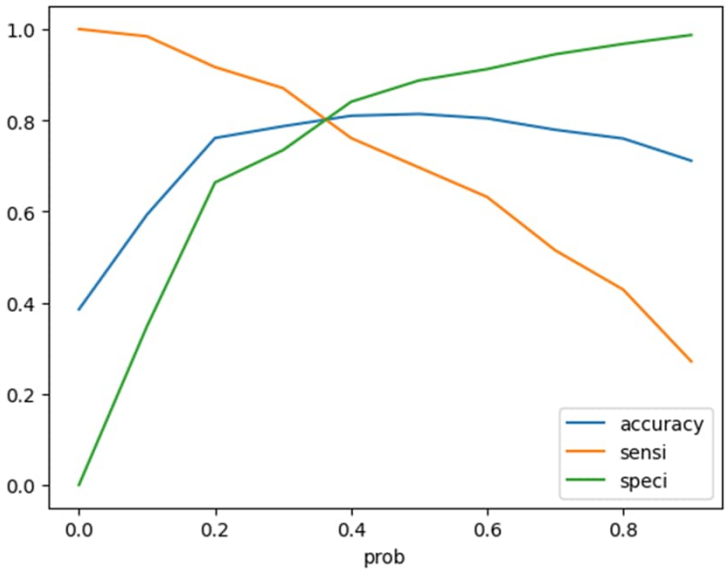
- Dummy variables created for object variables
  - Data Count
    - Column : 9074
    - Rows: 88 rows

- Model Building:
  - Train Test Split : 70 : 30 ratio
  - Feature Scaling : Numerical variables ('TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website')
  - Feature selection using RFE : select 15 variables as output
  - Building Model by removing the variable with P- value > 0.05 and vif value > 5
  - Run the model on test data

	Train	Test
Accuracy	80%	81%
Sensitivity	79%	79%
Specificity	81%	82%

ROC Curve:





Optimal cut off point: 0.4

Conclusion

Based on the model output we can conclude topmost variables which can be used by the company to target potential leads.

- Total Visits
- Total time spent on Website
- Last Notable Activity - Had a Phone conversation
- Lead add form from Lead origin
- What is your current occupation – Working Professional