1. **Which are the top three variables in your model which contribute most towards the probability of a lead getting converted?**

ANS:

The top three variables that contribute most towards the probability of a lead getting converted are as follows :

1. **Total Time Spent on website**
    a. Positive contribution
    b. Higher the time spent on the website, higher the probability of the lead converting into a customer
    c. Sales team should focus on such leads.
2. **Tags_Will revert after reading the email**
    a. Positive contribution
    b. If the tag chosen is 'will revert after reading the email', then there is a higher probability that the lead would convert.
3. **Leads Quality_Not Sure**
    a. Positive contribution

    b. If the lead Quality is 'not sure' , then there is a higher probability that the lead would convert.

2. **What are the top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion?**

ANS:

The top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion are:

1. Tags_Will revert after reading the email
2. Leads Quality_Not Sure

3. Last Notable Activity_SMS Sent

3. **X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as many of such people as possible. Suggest a good strategy they should employ at this stage.**

ANS:

1. Total time spent on Website
   we can say that Users spending more time on the website are more likely to get converted
2. "TotalVisits", "Page Views Per Visit" showing similar relation
   However they might be repeatedly visiting to compare courses from other websites and the number of visits are also more for that reason.
   We can say that Users visiting more time on website are more likely to get converted.
3. "What is your current occupation"

Working professionals have high conversion rate.To increase overall conversion rate, we need to increase the number of Working Professional leads by reaching out to them through different social sites such as LinkedIn etc. and also on increasing the conversion rate of Unemployed leads.
Students can be approached because after education they may preparing for getting jobs in industry.

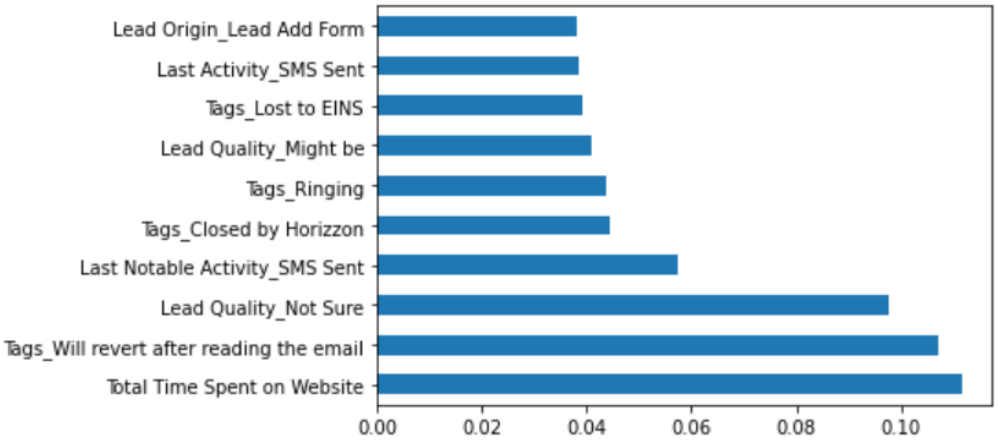
4. **Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.**
   ANS:
   1. In **Lead Origin**, Focus on **Lead Add Form** , their conversion is higher than other features.
   2. Focus on **working professionals** because their conversion ratio is higher.And do not focus on **'unemployed'** and 'students' because they might not purchase the course.
   3. In **Lead Source column**, **'google'** and 'Welingak' website and Reference has a higher rate of conversion than Olark chat, Organic Search, direct traffic
   4. In the **city** column, **Mumbai** region 's people have a higher rate of conversion than any other city.
   5. In **Last Notable Activity**, focus on only Olark chat Conversation **SMS Sent** because of their high conversion. And do not focus on modified and email opened as their low conversion rate.

==========

==========

1. **Which are the top three variables in your model which contribute most towards the probability of a lead getting converted?**

```
ranked_features=pd.Series(model.feature_importances_,index=X.columns)
ranked_features.nlargest(10).plot(kind='barh')
plt.show()
```



2. **What are the top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion?**

```
mutual_data=pd.Series(mutual_info,index=X.columns)
mutual_data.sort_values(ascending=False)
```

```
Total Time Spent on Website                 0.119649
Lead Quality_Not Sure                       0.101449
Tags_Ringing                                0.060144
Tags_Will revert after reading the email    0.060100
Last Notable Activity_SMS Sent              0.060025
                                               ...
Tags_wrong number given                     0.000000
Lead Source_Referral Sites                  0.000000
Lead Source_Other_Lead_Source               0.000000
Lead Source_Organic Search                  0.000000
Tags_invalid number                         0.000000
Length: 84, dtype: float64
```

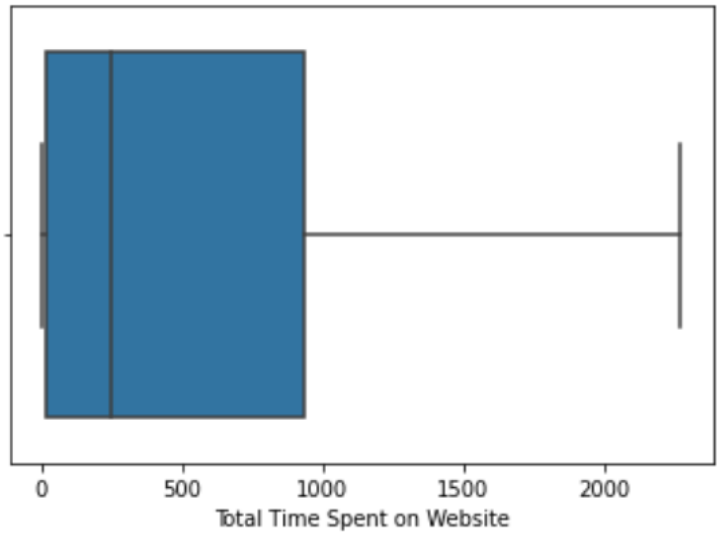
3. **X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.**
ans:

1. **Total time spent on Website**

Total Time Spent on Website

```
sns.boxplot(df["Total Time Spent on Website"])
```

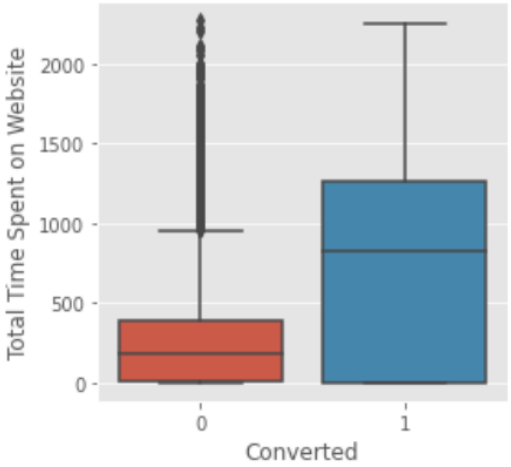
```
/usr/local/lib/python3.8/dist-packages/seaborn/_decorators.py:36:
  warnings.warn(
<matplotlib.axes._subplots.AxesSubplot at 0x7f8b15a08070>
```



plotted box plot on 'Total time spent on Website' column

```
fig=plt.subplots(figsize=(4, 4))
sns.boxplot(y = "Total Time Spent on Website", x = 'Converted', data = df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f549b9b5cd0>
```



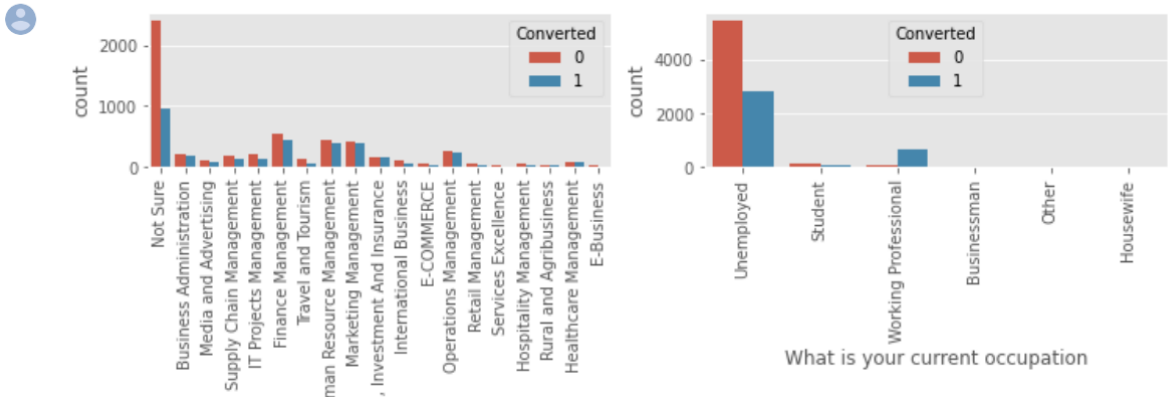
OBSERVATION:

The median of both the conversion and non-conversion differ a lot in Time spent on the website and hence we can say that Users spending more time on the website are more likely to get converted.

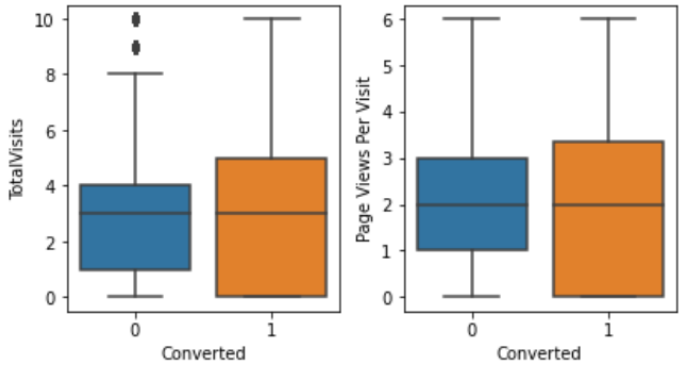
Websites can be made more appealing so as to increase the time of the Users on websites.

2. TotalVisits and Page views Per Visit

```
fig=plt.subplots(figsize=(10, 6))

for i, feature in enumerate(["Specialization", "What is your current occupation"]):
    plt.subplot(2, 2, i+1)
    plt.subplots_adjust(hspace = 2.0)
    sns.countplot(x=feature, hue="Converted", data=df)
    plt.xticks( rotation='vertical')
    plt.tight_layout()
```



```
fig=plt.subplots(figsize=(6, 6))

for i, feature in enumerate(["TotalVisits", "Page Views Per Visit"]):
    plt.subplot(2, 2, i+1)
    plt.subplots_adjust(hspace = 2.0)
    sns.boxplot(y = feature, x = 'Converted', data = df)
    plt.tight_layout()
```



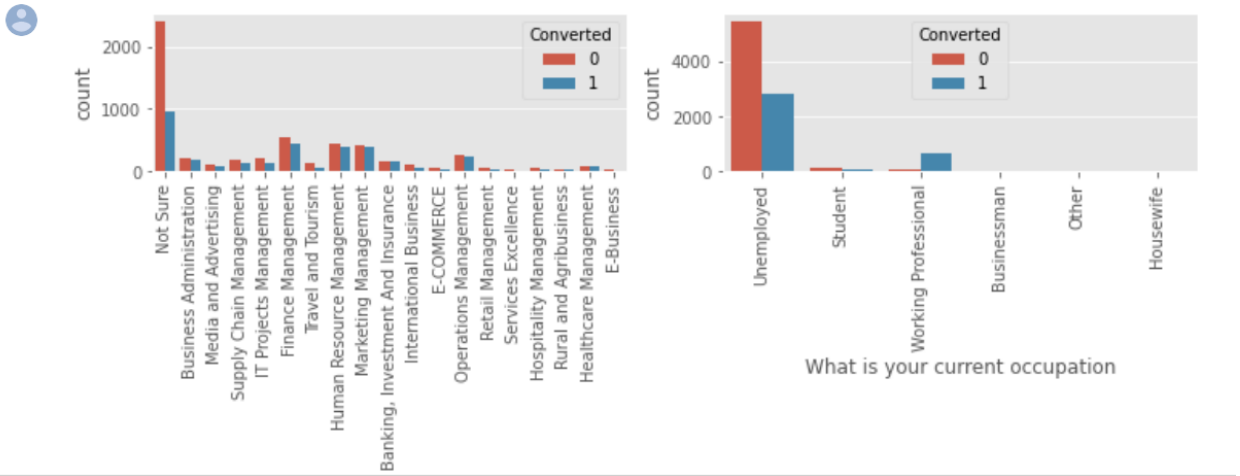
We check correlation of TotalVisits and Page Views Per Visit column
"TotalVisits", "Page Views Per Visit" showing similar relation to covered so we can say they have high multicollinearity between them so we can drop either of them.

```
[ ]  #will drop Page Views Per Visit
     df.drop(['Page Views Per Visit'], axis=1, inplace=True)
```

Observation:
"TotalVisits", "Page Views Per Visit" showing similar relation However they might be repeatedly visiting to compare courses from other websites and the number of visits are also more for that reason. We can say that Users visiting more time on a website are more likely to get converted.

3. “What is your current occupation”

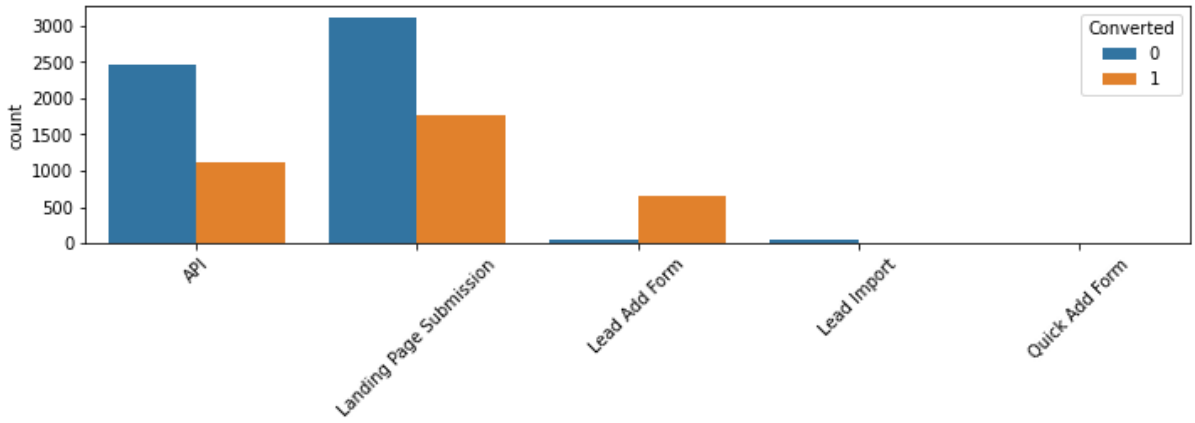
```
fig=plt.subplots(figsize=(10, 6))

for i, feature in enumerate(["Specialization", "What is your current occupation"]):
    plt.subplot(2, 2, i+1)
    plt.subplots_adjust(hspace = 2.0)
    sns.countplot(x=feature, hue="Converted", data=df)
    plt.xticks( rotation='vertical')
    plt.tight_layout()
```



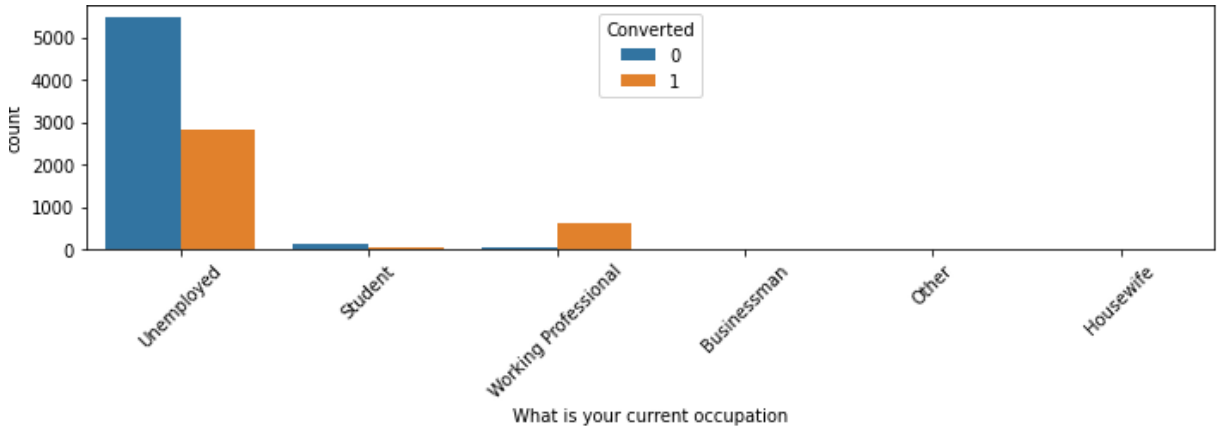
OBSERVATION:

Looking at above plot, no particular inference can be made for Specialization Looking at above plot, we can say that working professionals have high conversion rate Number of Unemployed leads are more than any other category To increase overall conversion rate, we need to increase the number of Working Professional leads by reaching out to them through different social sites such as LinkedIn etc. and also on increasing the conversion rate of Unemployed leads

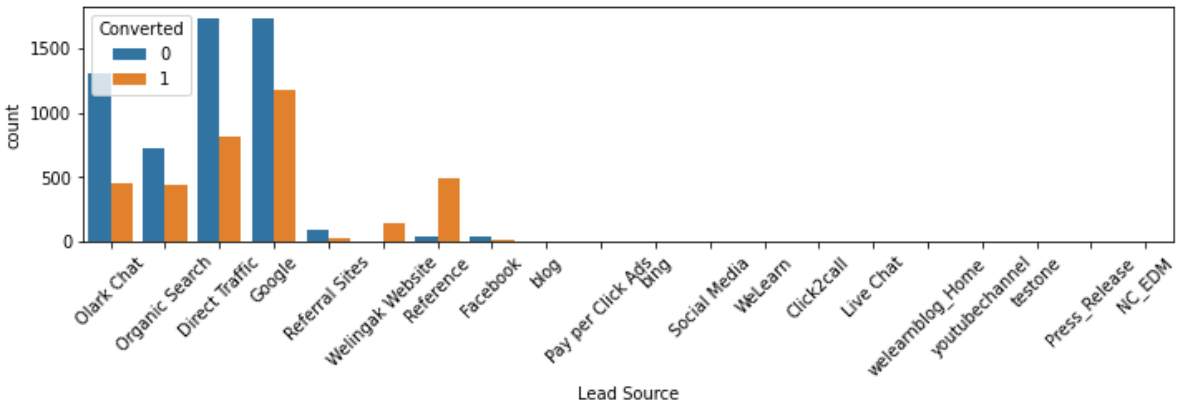
4. **Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.**



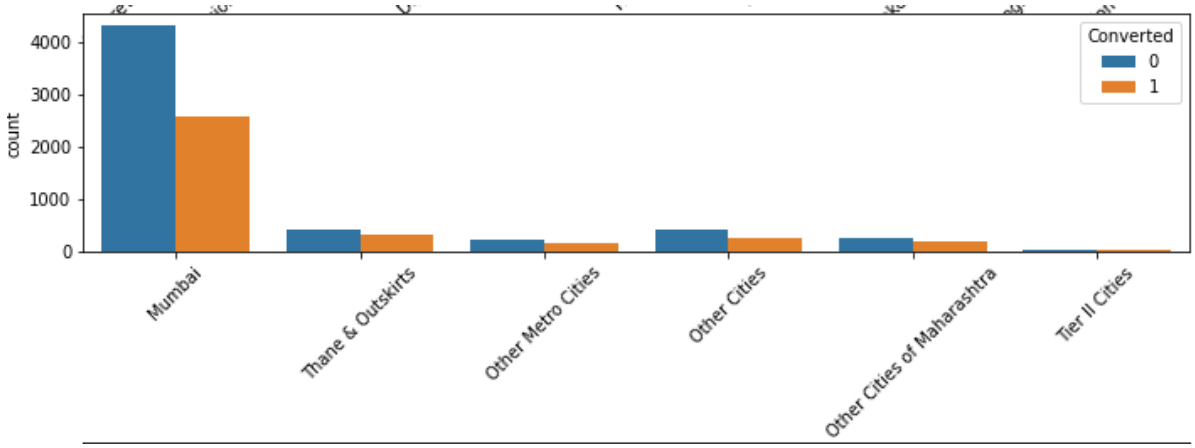
In **Lead Origin**, Focus on Lead Add Form , their conversion is higher than other features.



In this column, Focus on working professionals because their conversion ratio is higher. And do not focus on unemployed and students because they might not purchase the course.

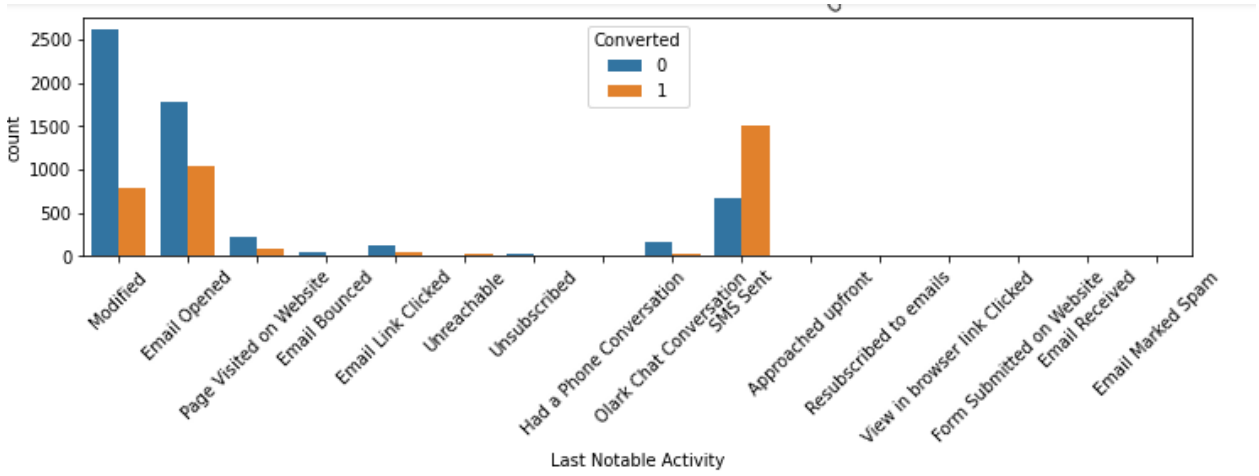


In **Lead Source column**, google and Welingak website and Reference has a higher rate of conversion than Olark chat, Organic Search, direct traffic



In the **city** column, Mumbai region ‘s people have a high rate of conversion than any other city.

ssssssss



In **Last Notable Activity**, focus on only Olark chat Conversation SMS Sent because of their high conversion. And do not focus on modified and email opened as their low conversion rate.

A Brief Summary

Problem Statement: This analysis is done for X Education and to find ways to get more industry professionals to join their courses.The basic data provided gave us a lot of information about how the potential customers visit the website, the time they spend there, how they reached the site and conversion rate.

The following are the steps used for achieving the model :

1. **Cleaning The Data :**

   Data cleaning is a process used to determine inaccurate, incomplete, or unreasonable data and then improve quality by correcting detected errors and omissions.First we converted ‘select’ value into nan because select is an option as not selected.

   Here, we first checked for the null values and dropped those columns where the null value percentage was greater than 70%.

   Then we dropped some columns which had null values greater than 45% which were not important.These columns included 'Asymmetrique Activity Index', 'Asymmetrique Profile Score', 'Asymmetrique Activity Score' and 'Asymmetrique Profile Index'.

   We also checked for the unique values of each column.
   We changed the nan values of columns into the frequent value.
   As the null value of ‘specialization’ column into ‘not sure’ because ‘not sure’ is occurring most frequently.
   In the same way we changed
   nan values of ‘city’ column into ‘Mumbai’ , nan values of ‘tags’ column into ‘Will revert after reading the email’ , nan value of ‘What matters most to you in choosing a course’ into ‘Better Career Prospects’ ,nan value of ’What is your current occupation’ into ‘Unemployed’,
   nan value of ‘Country’ column into ‘India’ and nan value of ‘Lead Quality’ column into ‘Not Sure’.
   We then replaced the null values of leftover columns with reference to central tendency as their null count is very less.
   As the ‘Lead source’ column has a value as null and ‘google’ so we replaced it with ‘Google’.
   In the same way we changed the null value of ‘TotalVisits’ column to mean value of ‘TotalVisits’ ,‘Page Views Per Visit’column to mean value of ‘Page Views Per Visit’and ‘Last Activity’column to ‘Email Opened’.
   Lastly, 'Prospect Id’ and 'Lead Number' are dropped because we don't require these columns for EDA.

2. **EDA:**

   Exploratory Data Analysis is a process of examining or understanding the data and extracting insights or main characteristics of the data.We can find that some numerical variables consist of very high values as compared to their respective means. That's why we have created charts using boxplot to understand the patterns. We have observed that the outliers are very high and we need to treat it. That was the reason we have retained 96%

quantile of the data and removed the max value from it. We also converted low count values into one variable.

3. **Dummy Variables:**
   Dummy variables were created to identify the categorical variables to convert. The variables were not numerical so we needed it to be converted into numeric values to build the logistic regression model.
   Now the data is numeric so we have split the data set into train and test data frames 80 % and 20 % Respectively.StandardScaler used for further scaling.
4. **Model Building :**
   Now all the data is numeric we can create a logistic regression model.Accuracy obtained was 91.7%. We checked the correlation between the columns and dropped those columns having high collinearity.Then we performed PCA(Principal Component Analysis) for dimensionality reduction.The accuracy obtained after performing PCA is 87.28%.

   .

5. **Dimension reduction:**

   Dimensionality reduction is a machine learning (ML) or statistical technique of reducing the amount of random variables in a problem by obtaining a set of principal variables.

   In our model we have reduced the number of features to 10 and received accuracy of 87.22%.