X EDUCATION-LEAD SCORING CASE STUDY

By : Mohammed Zakir

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. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%

The CEO has set a target lead conversion rate of 80%. The lead scoring model should help the sales team to prioritize potential leads that have a higher conversion chance and enable them to focus on communicating with them.

Goals of the Case Study:

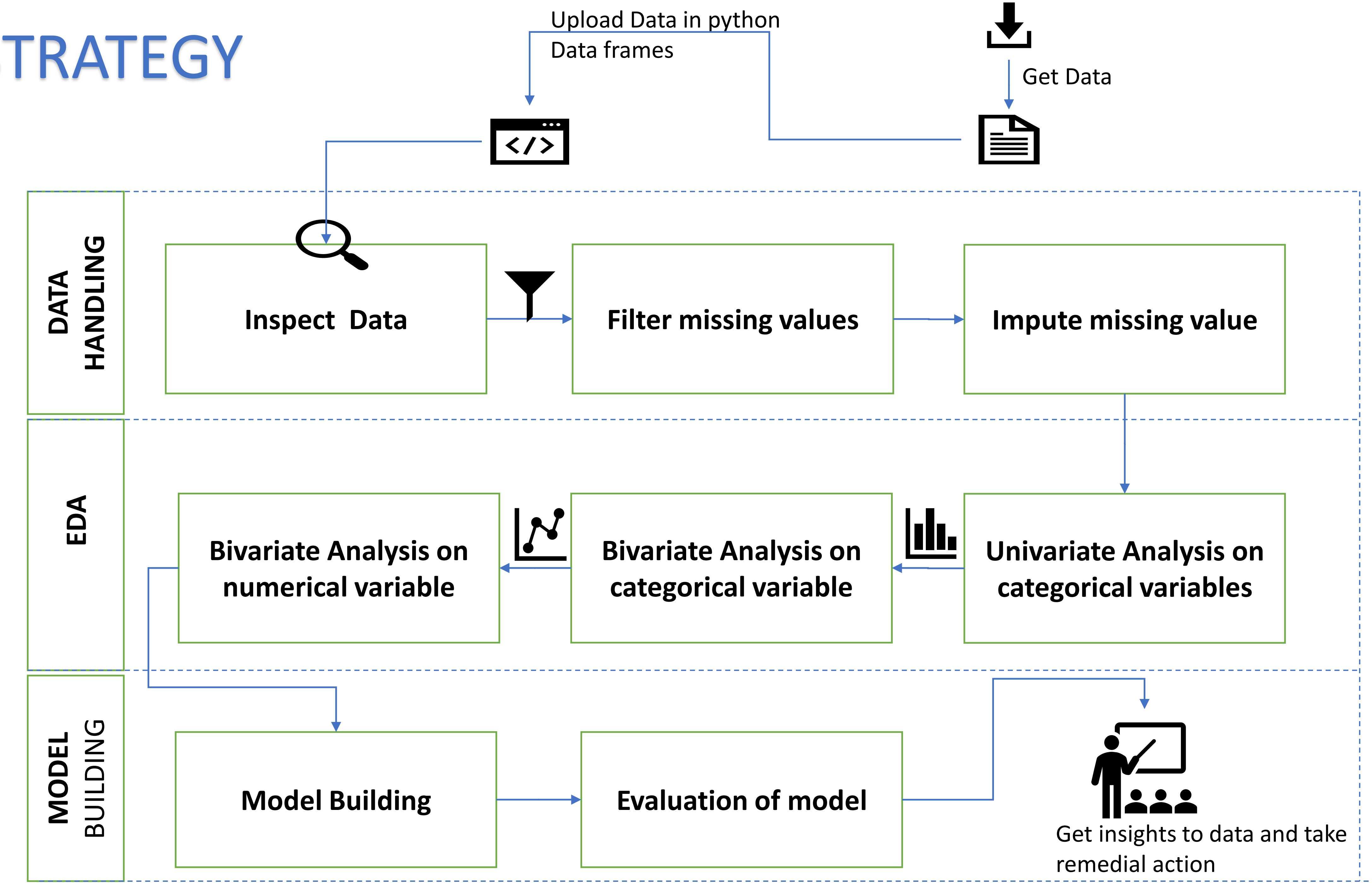
1.Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

2.There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so need to handle these as well

Content

- Strategy
- Understand Data
- Data Handling
- Univariate Analysis
- Bivariate Analysis
- Correlation Analysis
- Model Building
- Final Model
- Conclusion
- Recommendation

STRATEGY



Upload Data in python Data frames

Get Data

DATA HANDLING

Inspect Data

Filter missing values

Impute missing value

EDA

Univariate Analysis on categorical variables

Bivariate Analysis on categorical variable

Bivariate Analysis on numerical variable

MODEL BUILDING

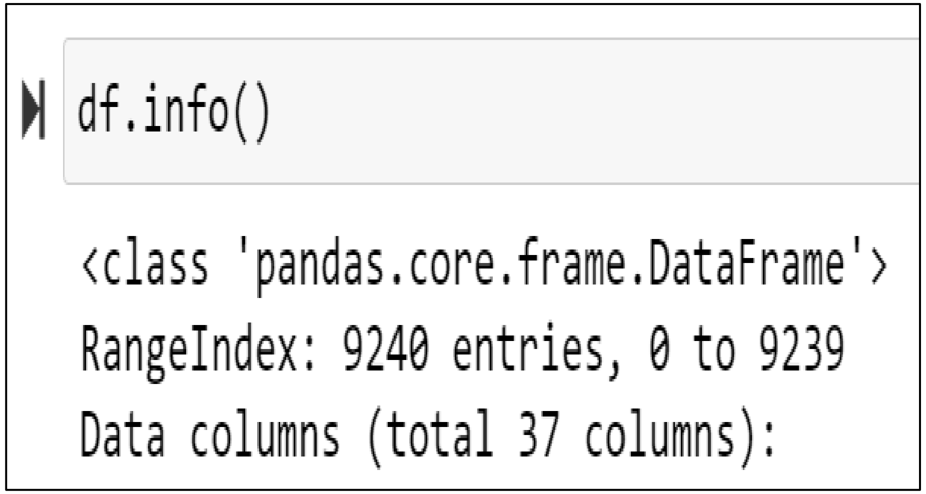
Model Building

Evaluation of model

Get insights to data and take remedial action

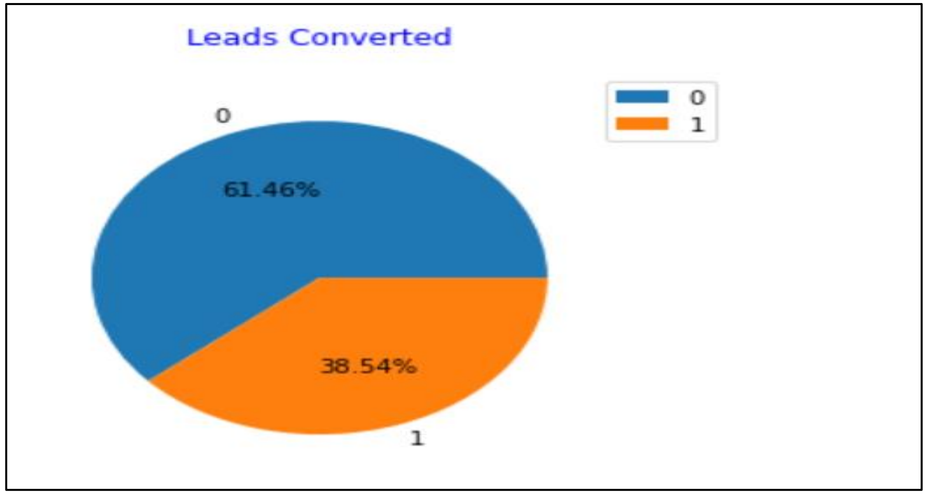
Understand Data

- Dataset used: **Leads.csv**
- Total Data point entries: **9240**
  - Number of column:**37**
  - Target column: **Converted**

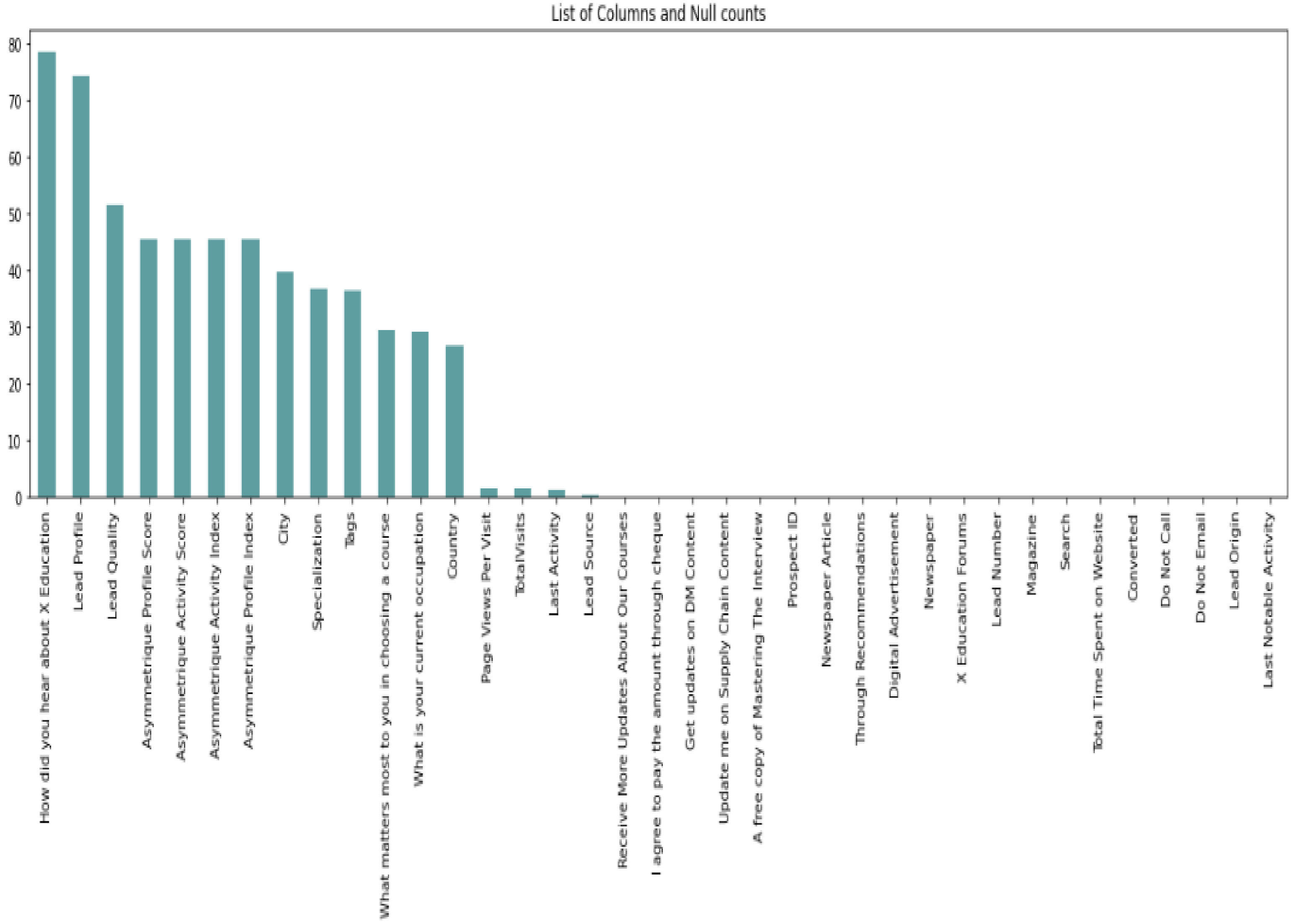


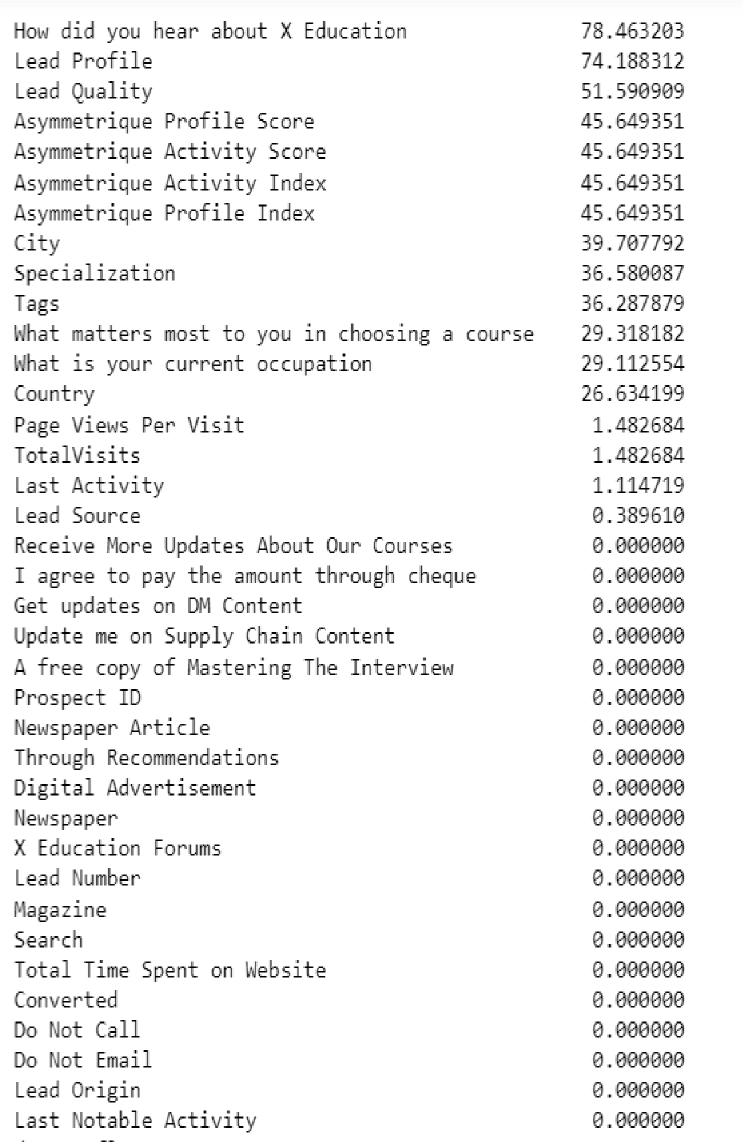
```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
```

- Current Conversion Rate:**38.54%**
  - Values in converted is binary
  - 0 means lead not converted
  - 1 means lead converted



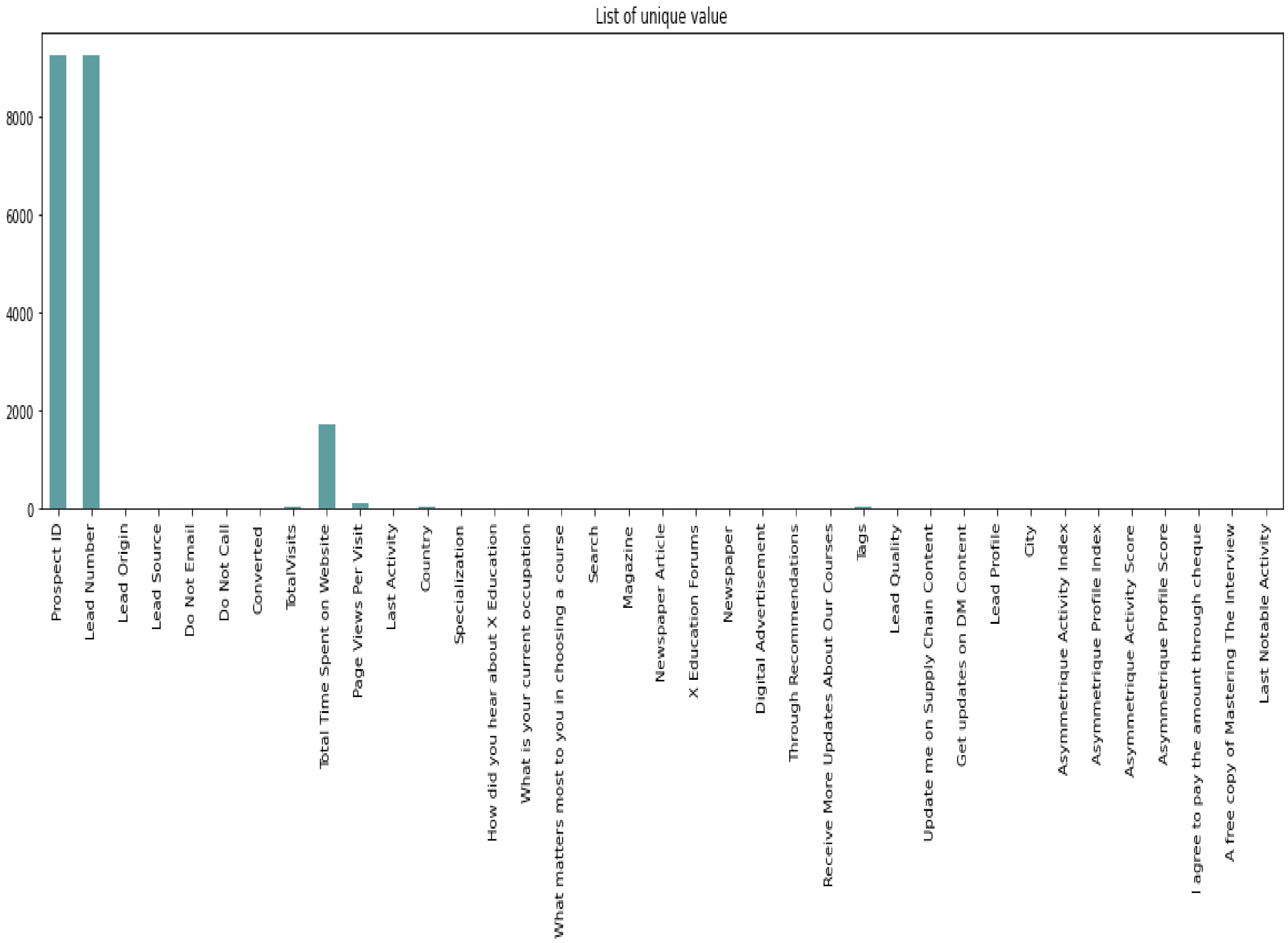
Data Preparation : Data Handling

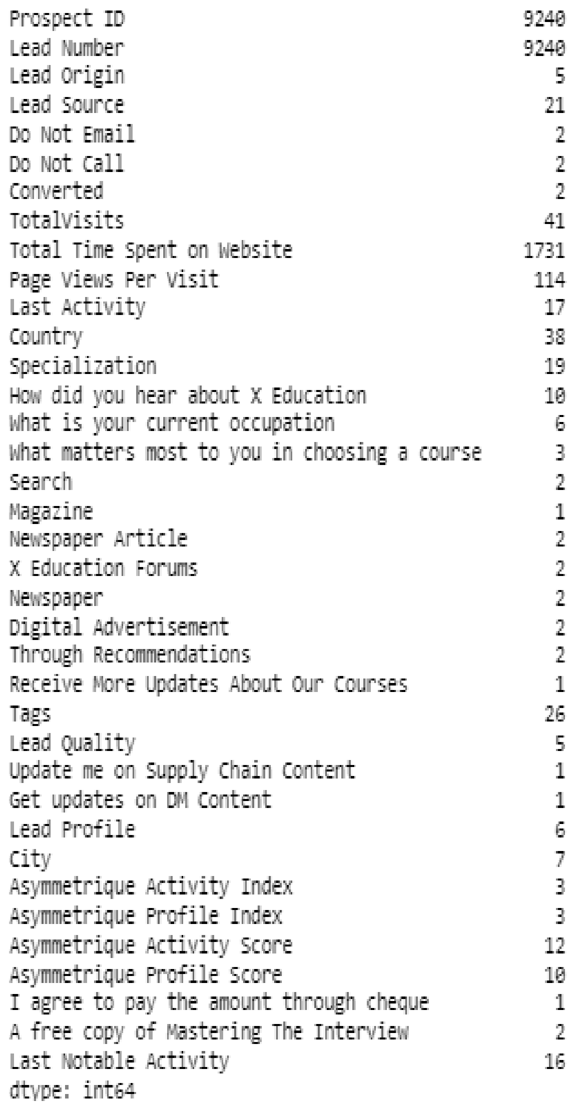




```
How did you hear about X Education          78.463203
Lead Profile                                74.188312
Lead Quality                                51.590909
Asymmetrique Profile Score                  45.649351
Asymmetrique Activity Score                 45.649351
Asymmetrique Activity Index                 45.649351
Asymmetrique Profile Index                  45.649351
City                                        39.707792
Specialization                              36.580087
Tags                                        36.287879
What matters most to you in choosing a course  29.318182
What is your current occupation             29.112554
Country                                     26.634199
Page Views Per Visit                         1.482684
TotalVisits                                  1.482684
Last Activity                                1.114719
Lead Source                                  0.389610
Receive More Updates About Our Courses       0.000000
I agree to pay the amount through cheque     0.000000
Get updates on DM Content                    0.000000
Update me on Supply Chain Content            0.000000
A free copy of Mastering The Interview       0.000000
Prospect ID                                  0.000000
Newspaper Article                            0.000000
Through Recommendations                      0.000000
Digital Advertisement                        0.000000
Newspaper                                    0.000000
X Education Forums                           0.000000
Lead Number                                  0.000000
Magazine                                     0.000000
Search                                       0.000000
Total Time Spent on Website                  0.000000
Converted                                    0.000000
Do Not Call                                  0.000000
Do Not Email                                 0.000000
Lead Origin                                  0.000000
Last Notable Activity                        0.000000
```

Data Preparation : Data Handling



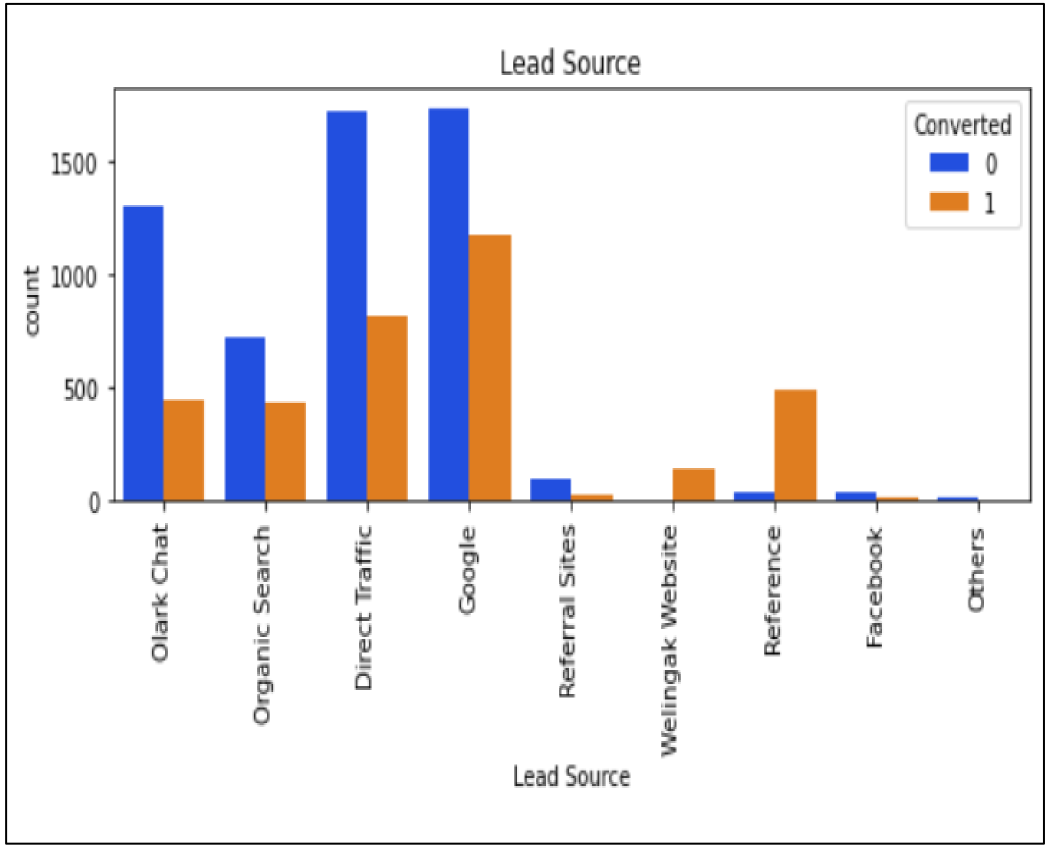


```
Prospect ID                                      9240
Lead Number                                      9240
Lead Origin                                         5
Lead Source                                        21
Do Not Email                                        2
Do Not Call                                         2
Converted                                           2
TotalVisits                                        41
Total Time Spent on Website                      1731
Page Views Per Visit                              114
Last Activity                                      17
Country                                            38
Specialization                                     19
How did you hear about X Education                 10
What is your current occupation                     6
What matters most to you in choosing a course       3
Search                                              2
Magazine                                            1
Newspaper Article                                   2
X Education Forums                                  2
Newspaper                                           2
Digital Advertisement                               2
Through Recommendations                             2
Receive More Updates About Our Courses              1
Tags                                               26
Lead Quality                                        5
Update me on Supply Chain Content                   1
Get updates on DM Content                           1
Lead Profile                                        6
City                                                7
Asymmetrique Activity Index                         3
Asymmetrique Profile Index                          3
Asymmetrique Activity Score                        12
Asymmetrique Profile Score                         10
I agree to pay the amount through cheque            1
A free copy of Mastering The Interview              2
Last Notable Activity                              16
dtype: int64
```

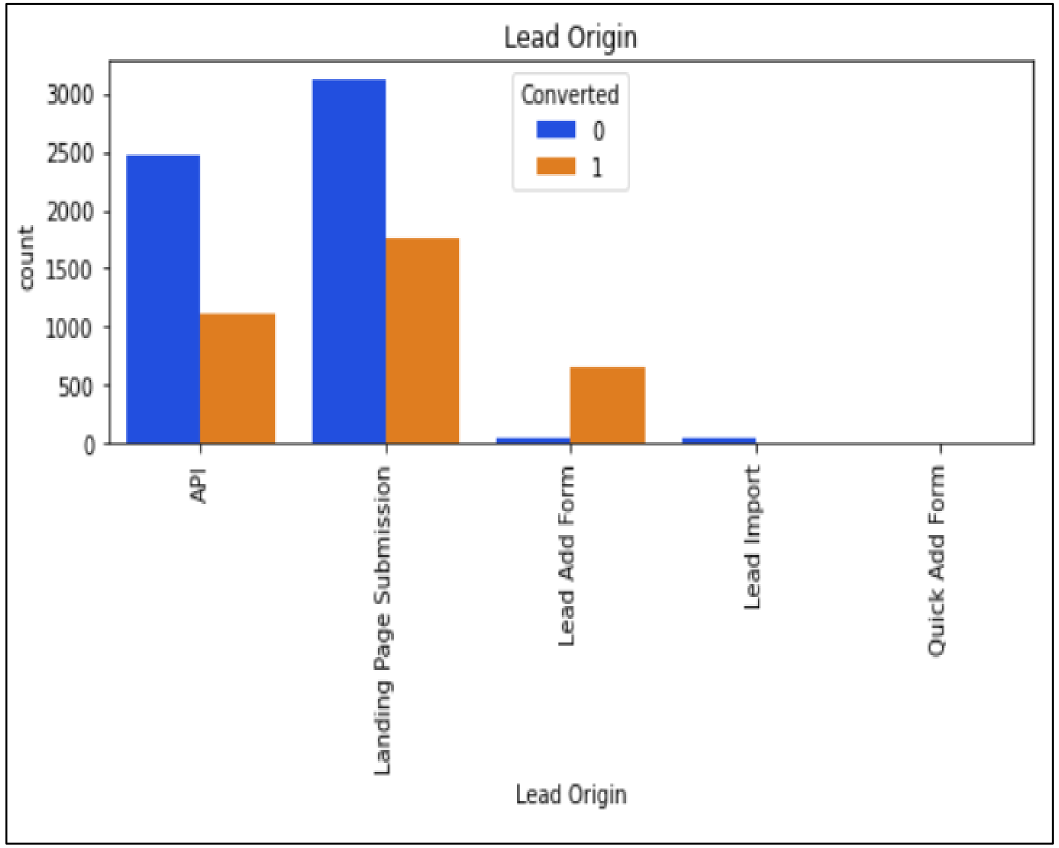
Data Preparation : Data Handling

- There are some categorical features having a label as “SELECT”.
  - This means the person might not have selected any value for that field. Hence this is as good as a missing value.
- Identifying all the missing data
  - Dropped columns having more than 40% null values
- Columns have only one category of response (unique) can be dropped
- Highly skewed column can also be dropped

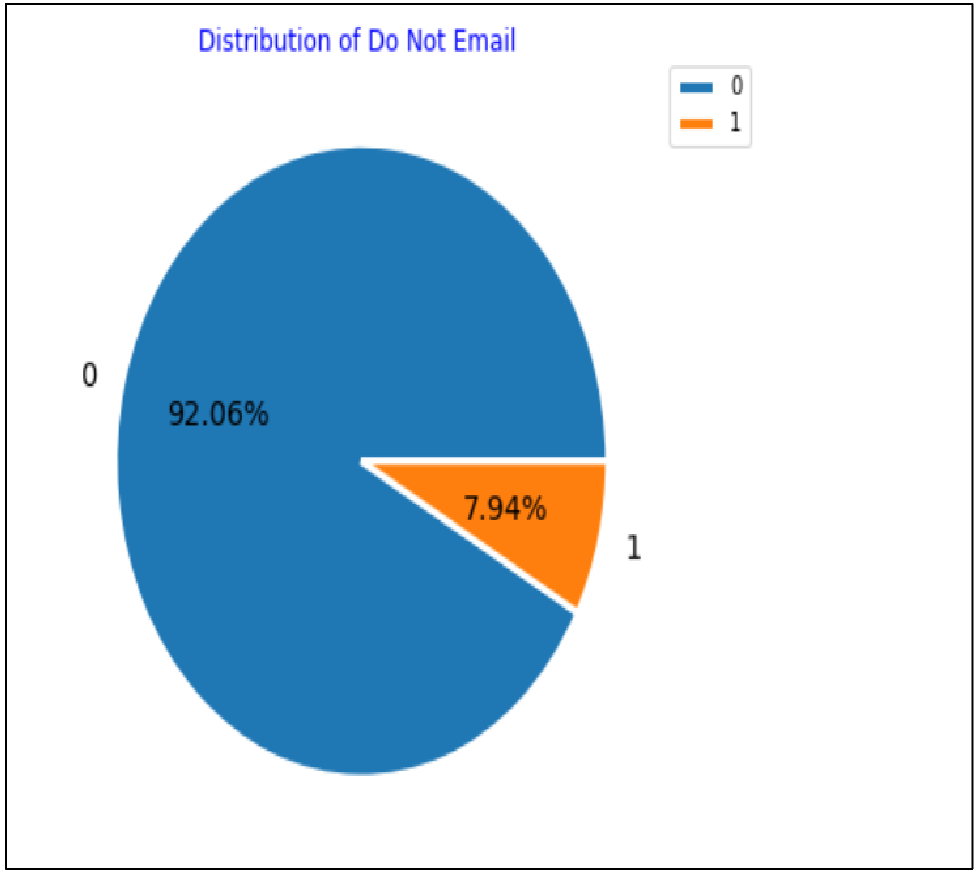
UNIVARIATE ANALYSIS



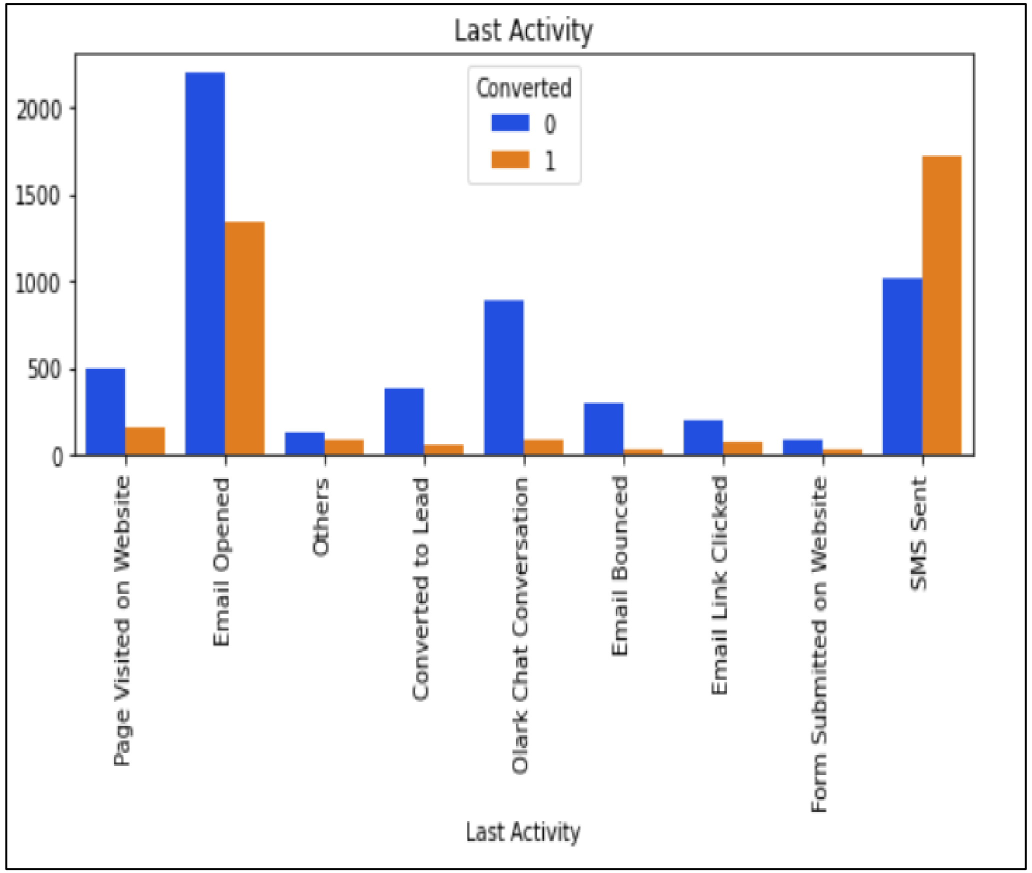
- The count of leads from the Google and Direct Traffic is maximum
- The conversion rate of the leads from Reference and Welingak Website is maximum



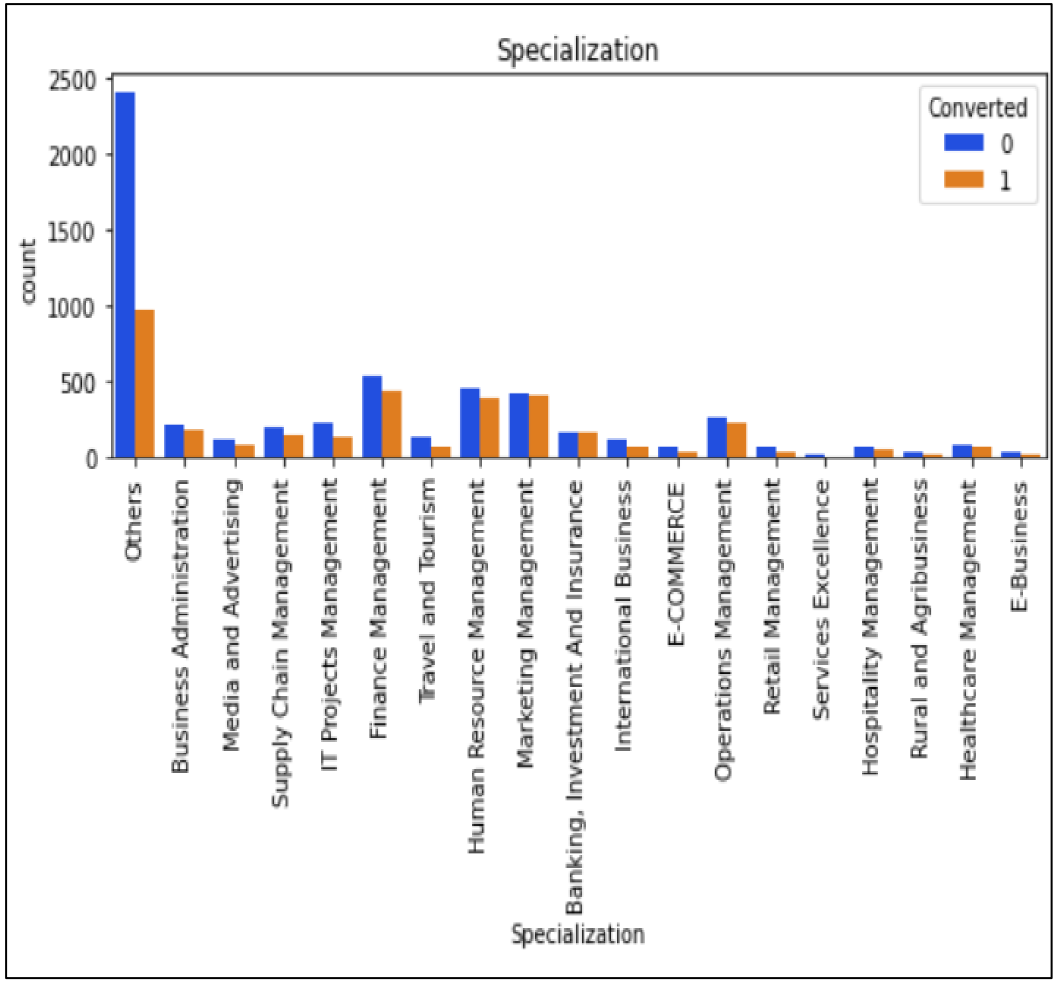
- API and Landing Page Submission has less conversion rate(~30%) but counts of the leads from them are considerable
- The count of leads from the Lead Add Form is pretty low but the conversion rate is very high
- Lead Import has very less count as well as conversion rate and hence can be ignored



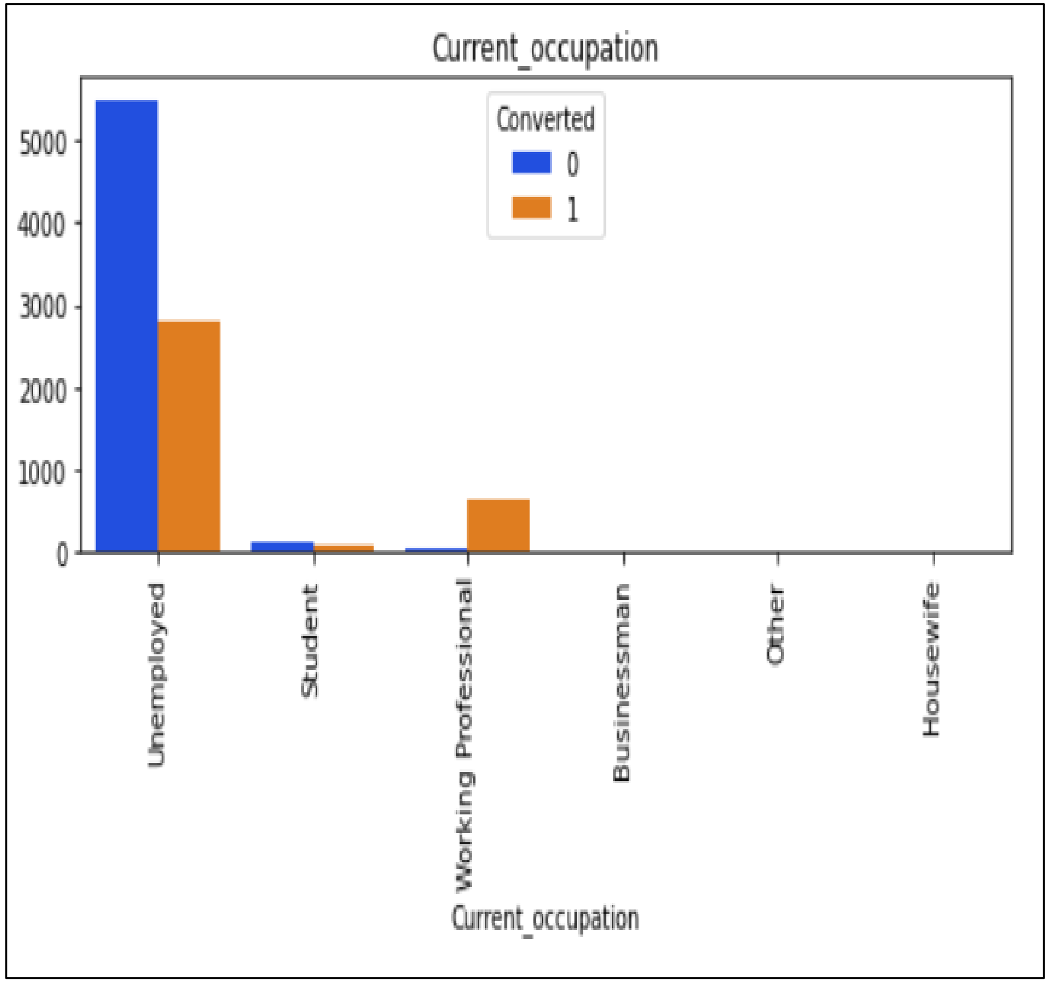
- A large proportion of customers, 92.1%, do not want to receive emails about the course.



- The count of activity as “Email Opened” is max
- The conversion rate of SMS sent as last activity is maximum

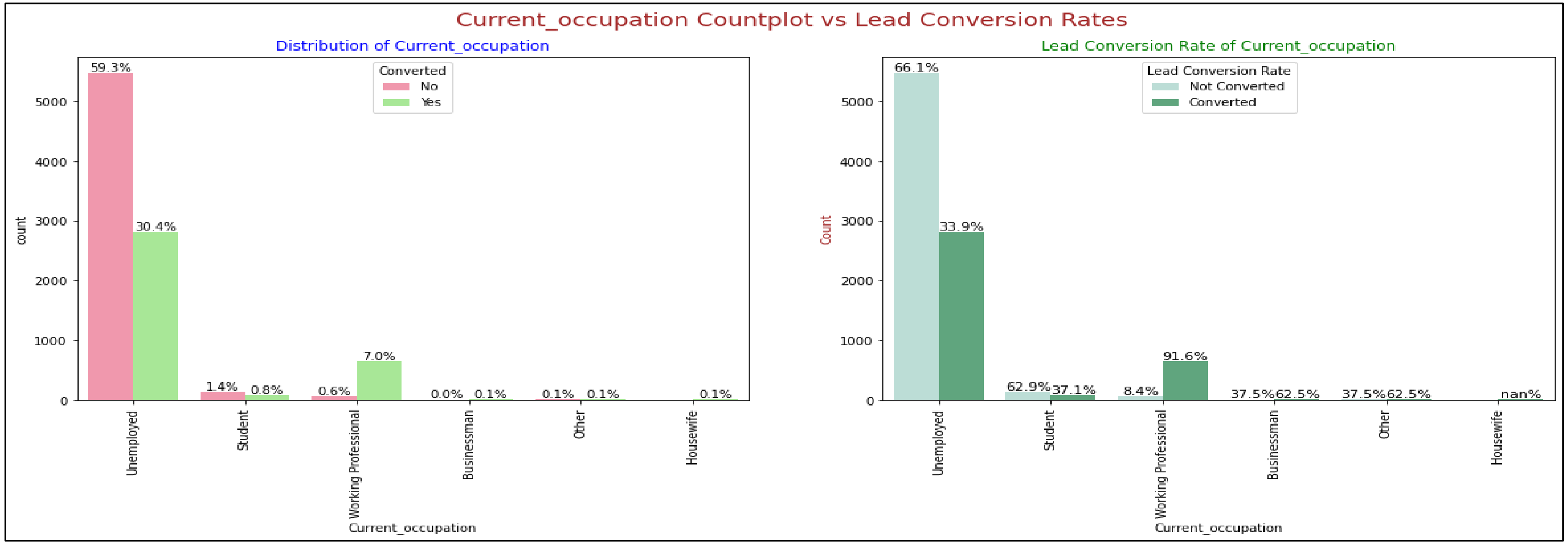


- Looking at above plot, no particular inference can be made for Specialization

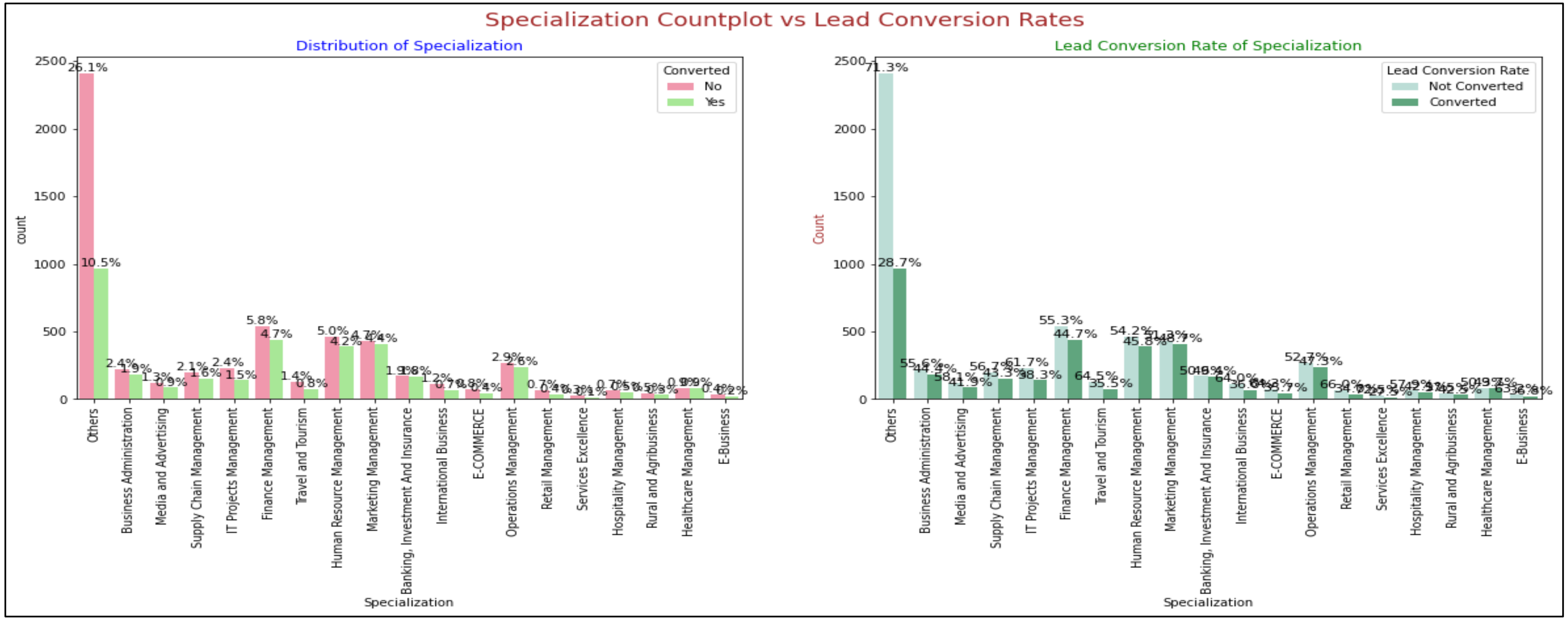


- Looking at above plot, we can say that working professionals have high conversion rate
- Number of Unemployed leads are more than any other category

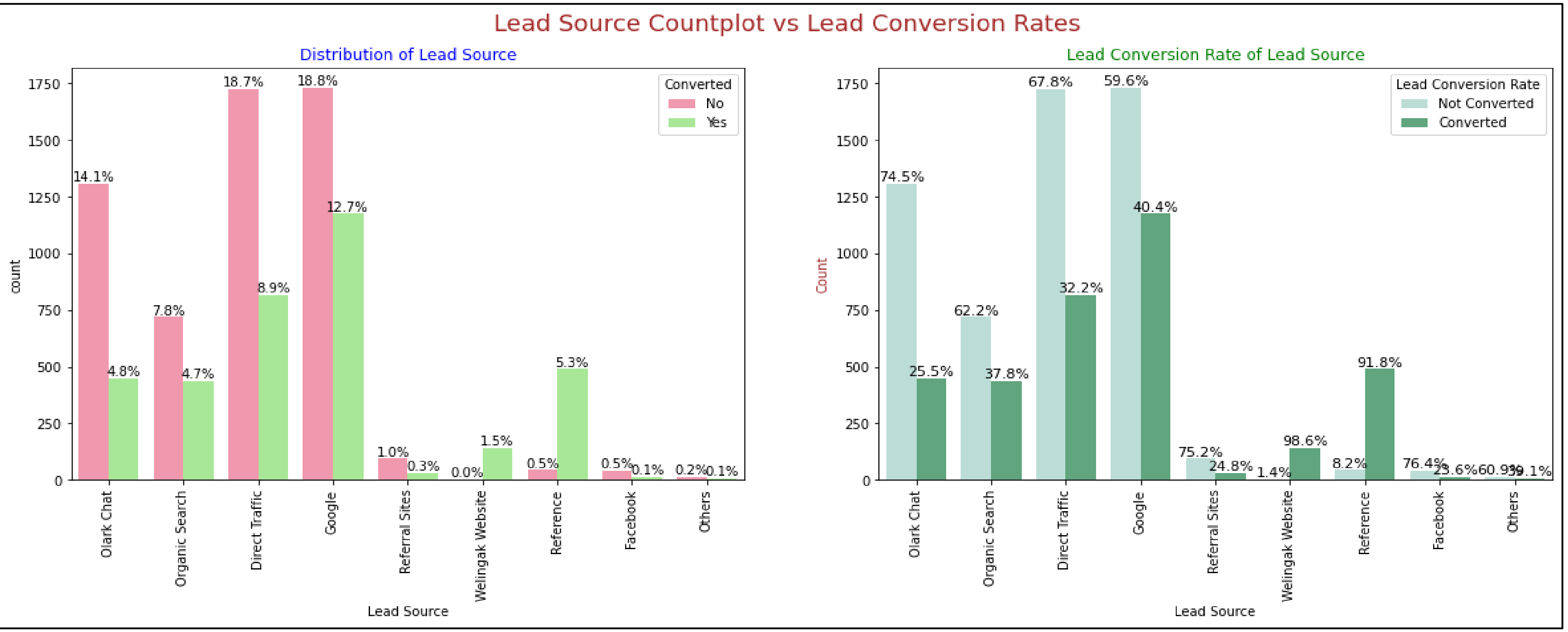
BIVARIATE ANALYSIS



Current_Occupation: Working Professionals have a significantly higher LCR at 91.6% compared to Unemployed people at 33.9%.



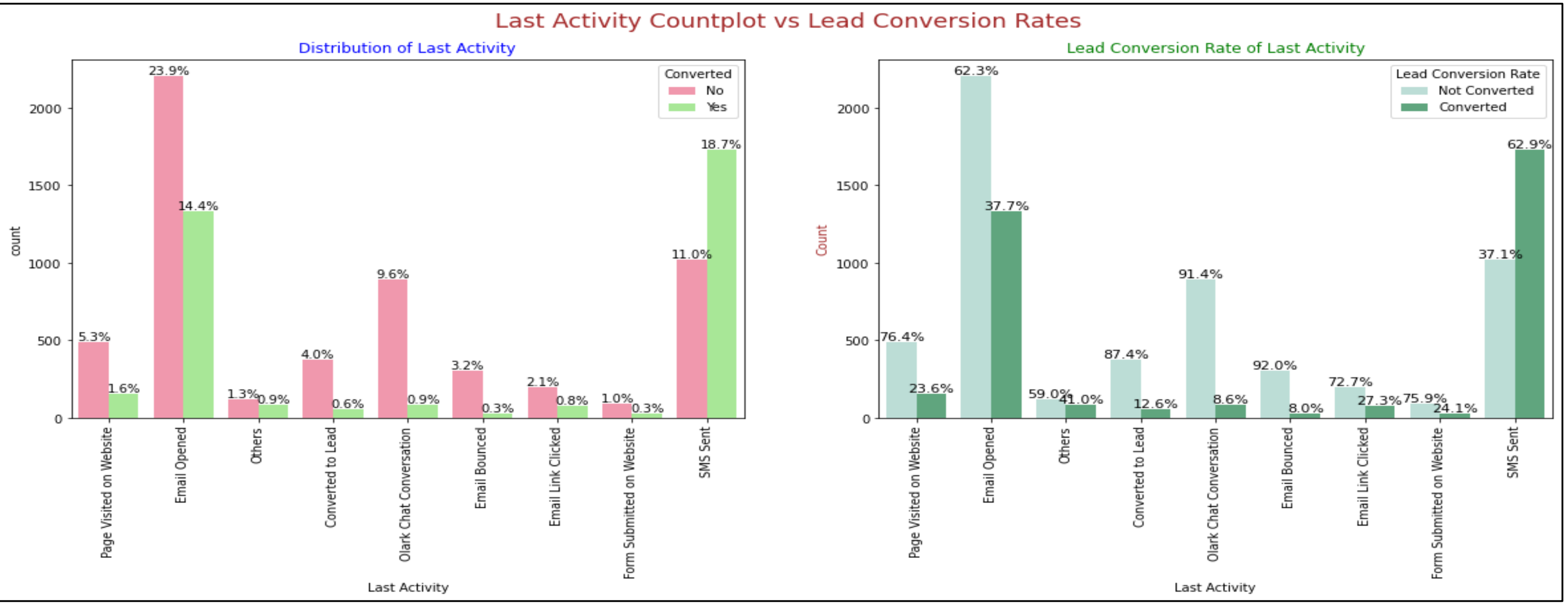
Specialization: Marketing Management, HR Management, Finance Management and Operations Management all show good LCRs, indicating a strong interest among customers in these specializations.


Lead Source Countplot vs Lead Conversion Rates

Distribution of Lead Source

Lead Conversion Rate of Lead Source

Lead Source: Google is the most effective Lead Source with an LCR of 40.4%

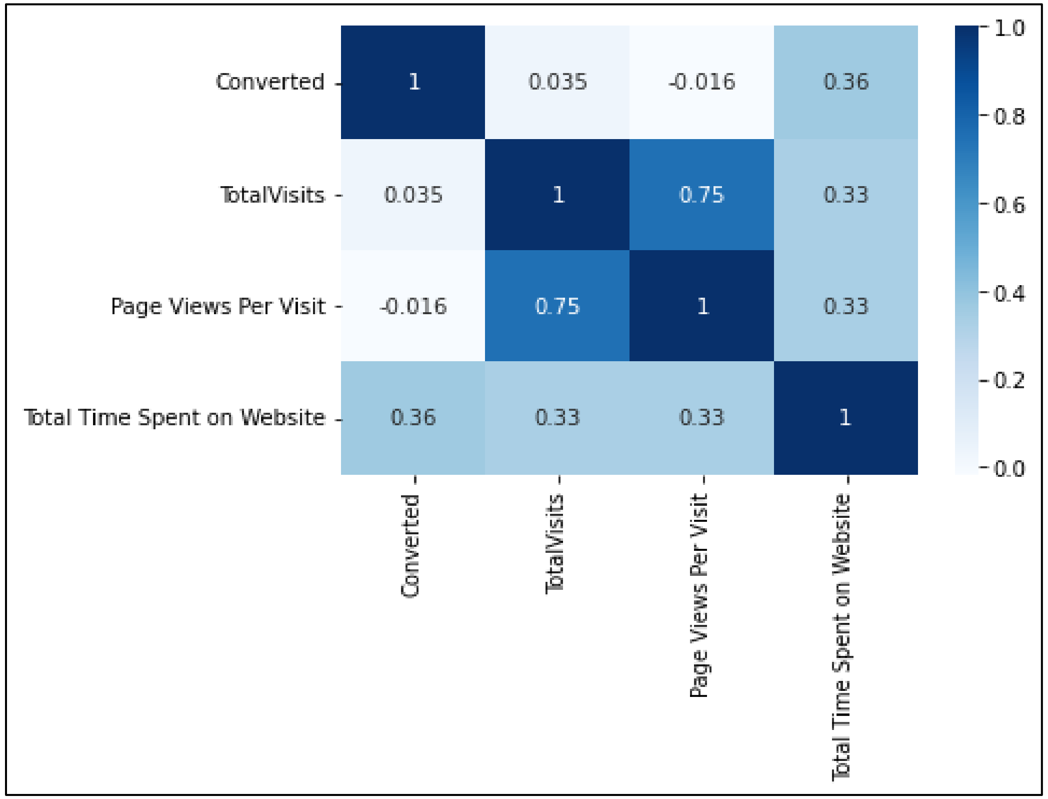

Last Activity Countplot vs Lead Conversion Rates

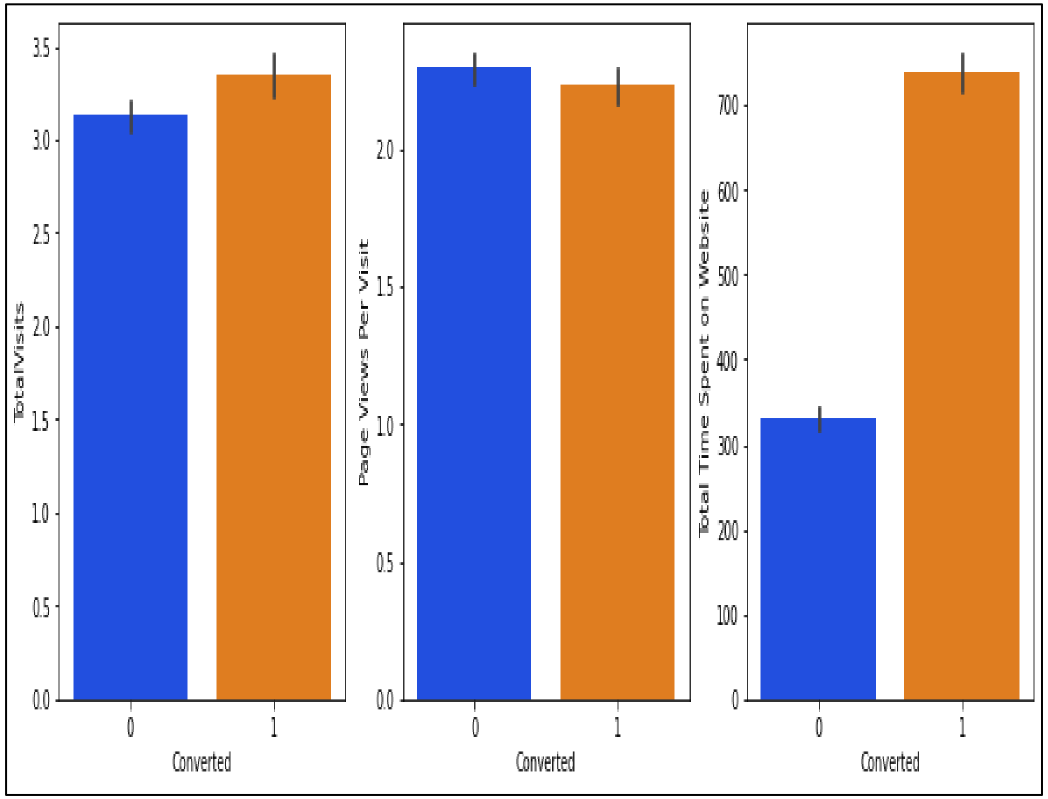
Distribution of Last Activity

Lead Conversion Rate of Last Activity

Last Activity: SMS Sent and Email Opened are the most effective Last Activity types with LCRs of 62.9% and 37.7% respectively.

CORRELATION ANALYSIS





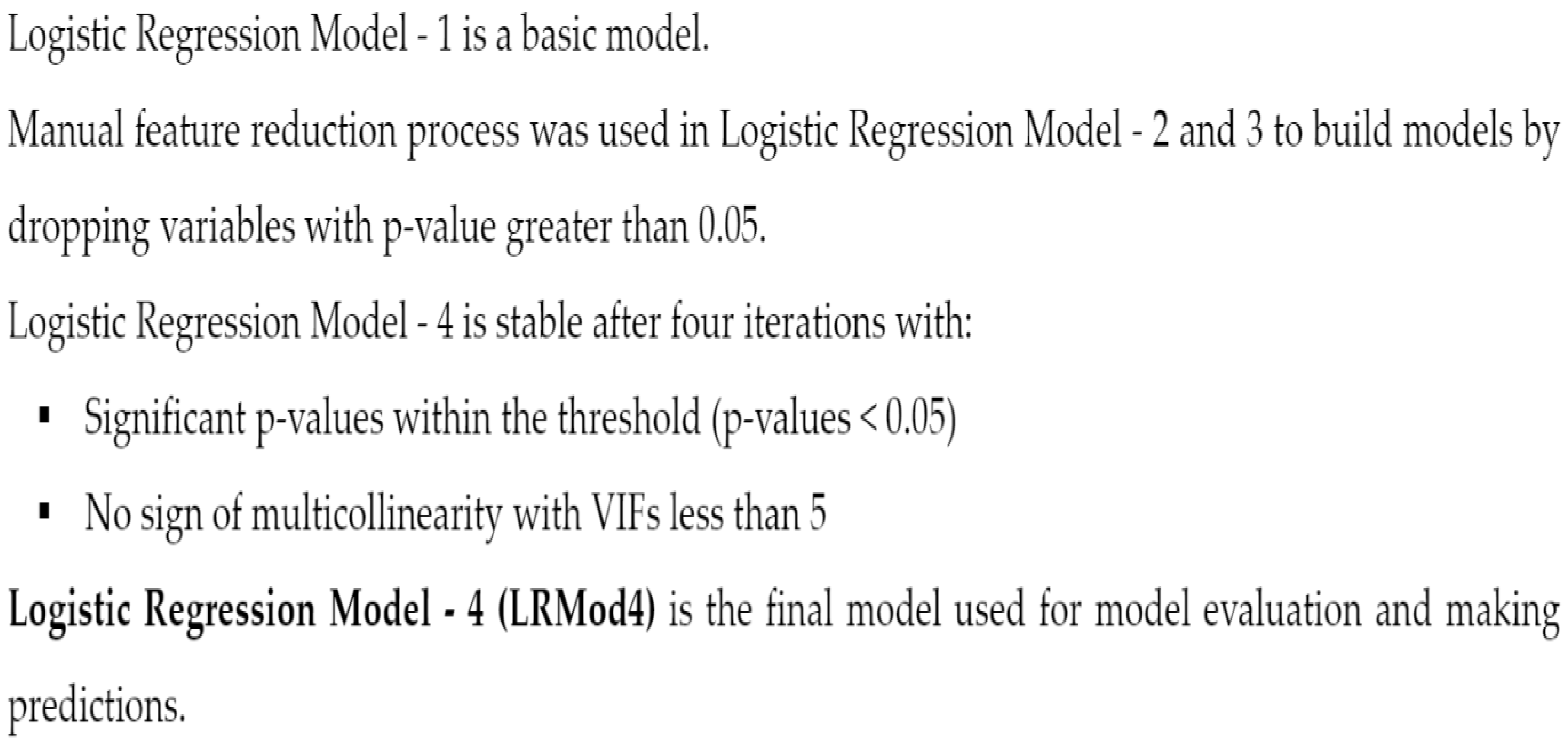
Inference:

- There is a strong positive correlation between 'Total Visits' and 'Page Views per Visit', indicating that customers who visit the website more frequently tend to view more pages per visit.
- Customers who spend more time on the website have a higher LCR, indicating that increasing the time spent on the website can lead to higher conversion rates.

MODEL BUILDING

- Dummy Variable Creation
  - As logistic regression can work with numeric data only, creating dummy variables for the categorical columns.
- Splitting Data into Training and Test set
  - Next, the dataset was split into training and test set, to train model first with a chunk of data and then evaluate its performance on unseen data.
- Feature Scaling
  - Feature Scaling is required before Logistic Regression to bring all the features in same scale, this ensures that features with high magnitude are not given higher importance by Logistic Regression Model.
- Model Building (Feature Selection Using RFE, Improvising the model further VIF and p-vales)
- Next, we are using RFE to obtain top 15 features to begin with.
- After that, manually inspecting p-values and VIF to improve the model even further.

Final Model



Logistic Regression Model - 1 is a basic model.

Manual feature reduction process was used in Logistic Regression Model - 2 and 3 to build models by dropping variables with p-value greater than 0.05.

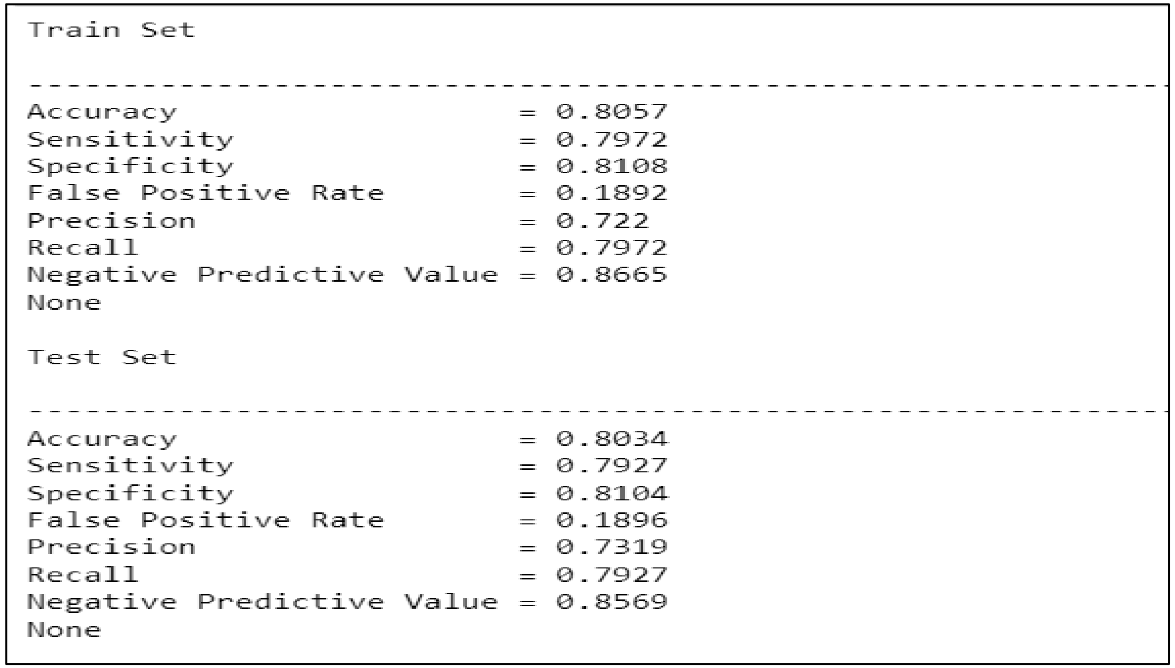
Logistic Regression Model - 4 is stable after four iterations with:

- Significant p-values within the threshold (p-values < 0.05)
- No sign of multicollinearity with VIFs less than 5

Logistic Regression Model - 4 (LRMod4) is the final model used for model evaluation and making predictions.

Model Evaluation

- Here, accuracy is the percentage of correctly predicted labels among all the labels
- Sensitivity = True Positive/(True Positive + False Negative)
  - Sensitivity gives us the percentage of correctly predicted conversion out of total conversions.
- Specificity = True Negative/(True Negative + False Positive)
  - Specificity gives us the percentage of correctly predicted non-conversion out of total non-conversions.



```
Train Set

-------------------------------------------------------------
Accuracy                  = 0.8057
Sensitivity               = 0.7972
Specificity               = 0.8108
False Positive Rate       = 0.1892
Precision                 = 0.722
Recall                    = 0.7972
Negative Predictive Value = 0.8665
None

Test Set

-------------------------------------------------------------
Accuracy                  = 0.8034
Sensitivity               = 0.7927
Specificity               = 0.8104
False Positive Rate       = 0.1896
Precision                 = 0.7319
Recall                    = 0.7927
Negative Predictive Value = 0.8569
None
```

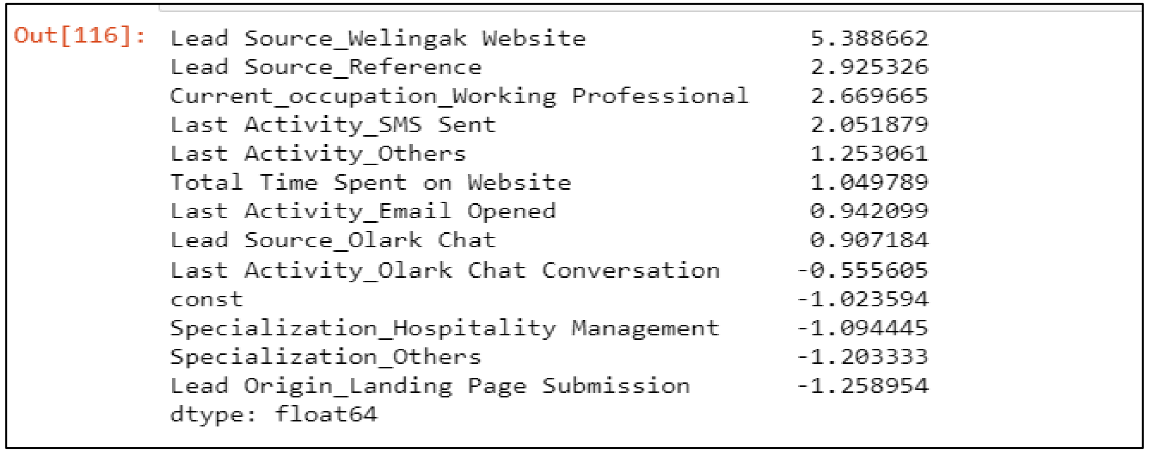
Final conclusion:

- As our model predicted leads from **Lead Source_Welingak Website, Lead Source_Reference, Current_occupation_Working Professional** are likely to be converted more, so should focus on these leads
- Company should focus on leads which our model predicted as 1 and likely focus attention to them as they are potential paying customer.
- Model predicted **last_activity_SMS_sent and Activity_Email_opened** are potentials leads and likely to convert.
  - Tailor made SMS or Emails to potentials leads to lure them and increase chance of conversion
- **Total Spent on Website** indicates that consumer checking out website are likely to convert as it implies interest shown by consumer.
  - Enabling customized Ads or contact information on website will help company to lure consumer and increase chance of conversion
- **Lead Source Olak chat** indicates consumers are showing interest and gathering information using chat feature.
- Getting customer details and reaching out to them may turn them to potential lead.

Recommendations

To minimize the rate of useless phone calls we would suggest:

- **Work on website and application** of company so customer can navigate and surf easily. Bad UI interface or badly designed website page are huge turn off for customer
- **Chatbot** to solve minor query of customer on website or app page is added plus
- **Customized SMS and emails** to customer based on their profiles/Bio
- **Referrals and incentives**-based scheme can spread word of mouth about company.
- Also **Marketing** and making consumer aware plays huge part to garner attention of consumer
- Reaching out to existing consumer timely and getting their **feedbacks** should be prioritized



```
Out[116]: Lead Source_Welingak Website                5.388662
          Lead Source_Reference                       2.925326
          Current_occupation_Working Professional     2.669665
          Last Activity_SMS Sent                      2.051879
          Last Activity_Others                        1.253061
          Total Time Spent on Website                 1.049789
          Last Activity_Email Opened                  0.942099
          Lead Source_Olark Chat                      0.907184
          Last Activity_Olark Chat Conversation      -0.555605
          const                                      -1.023594
          Specialization_Hospitality Management      -1.094445
          Specialization_Others                      -1.203333
          Lead Origin_Landing Page Submission        -1.258954
          dtype: float64
```

THANK YOU