Lead Scoring Case Study

Shreshth Verma

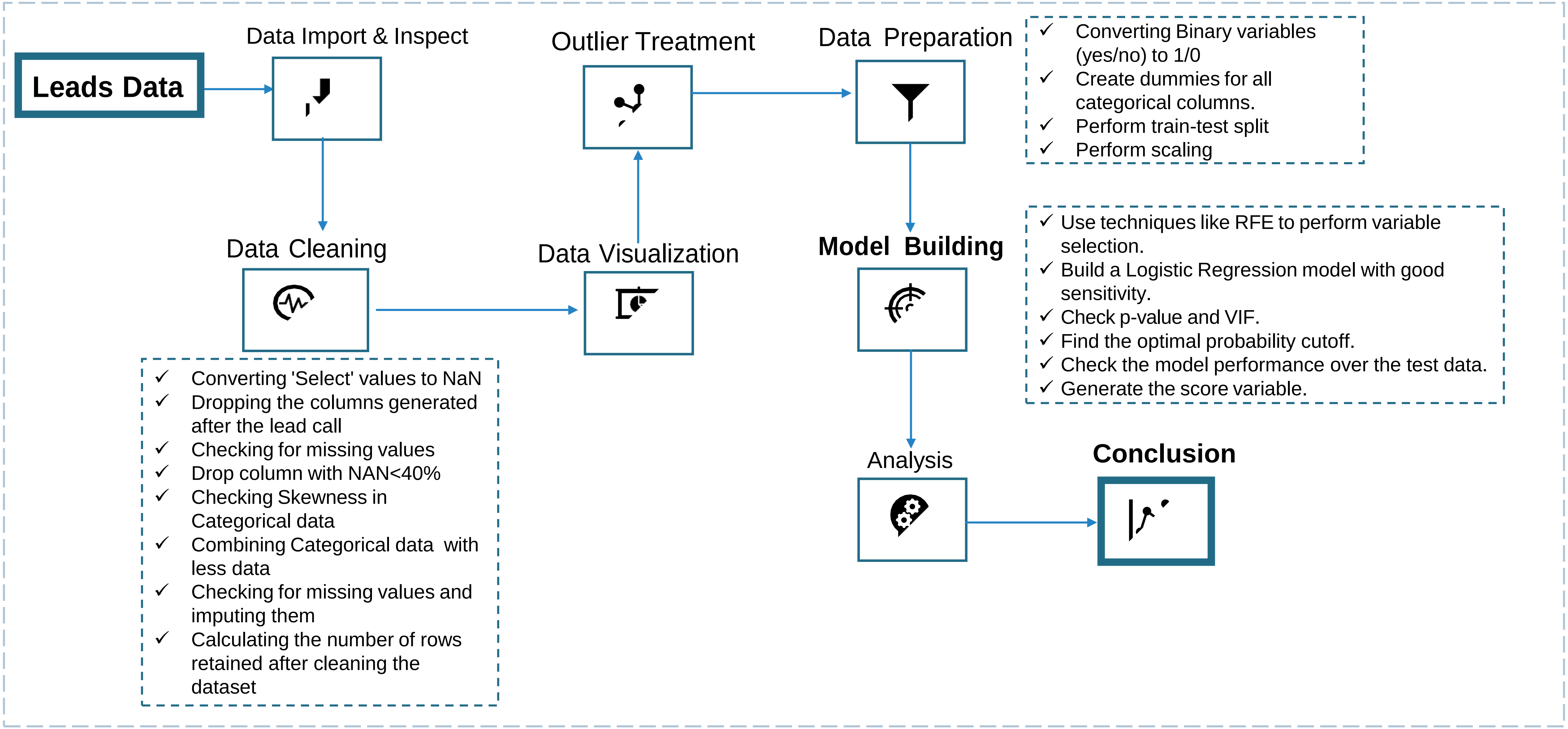
Business understanding

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. 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 company needs a model wherein a lead score is assigned to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

Data Workflow



Leads Data

Data Import & Inspect

Data Cleaning

- ✓ Converting 'Select' values to NaN
- ✓ Dropping the columns generated after the lead call
- ✓ Checking for missing values
- ✓ Drop column with NAN<40%
- ✓ Checking Skewness in Categorical data
- ✓ Combining Categorical data with less data
- ✓ Checking for missing values and imputing them
- ✓ Calculating the number of rows retained after cleaning the dataset

Data Visualization

Outlier Treatment

Data Preparation

- ✓ Converting Binary variables (yes/no) to 1/0
- ✓ Create dummies for all categorical columns.
- ✓ Perform train-test split
- ✓ Perform scaling

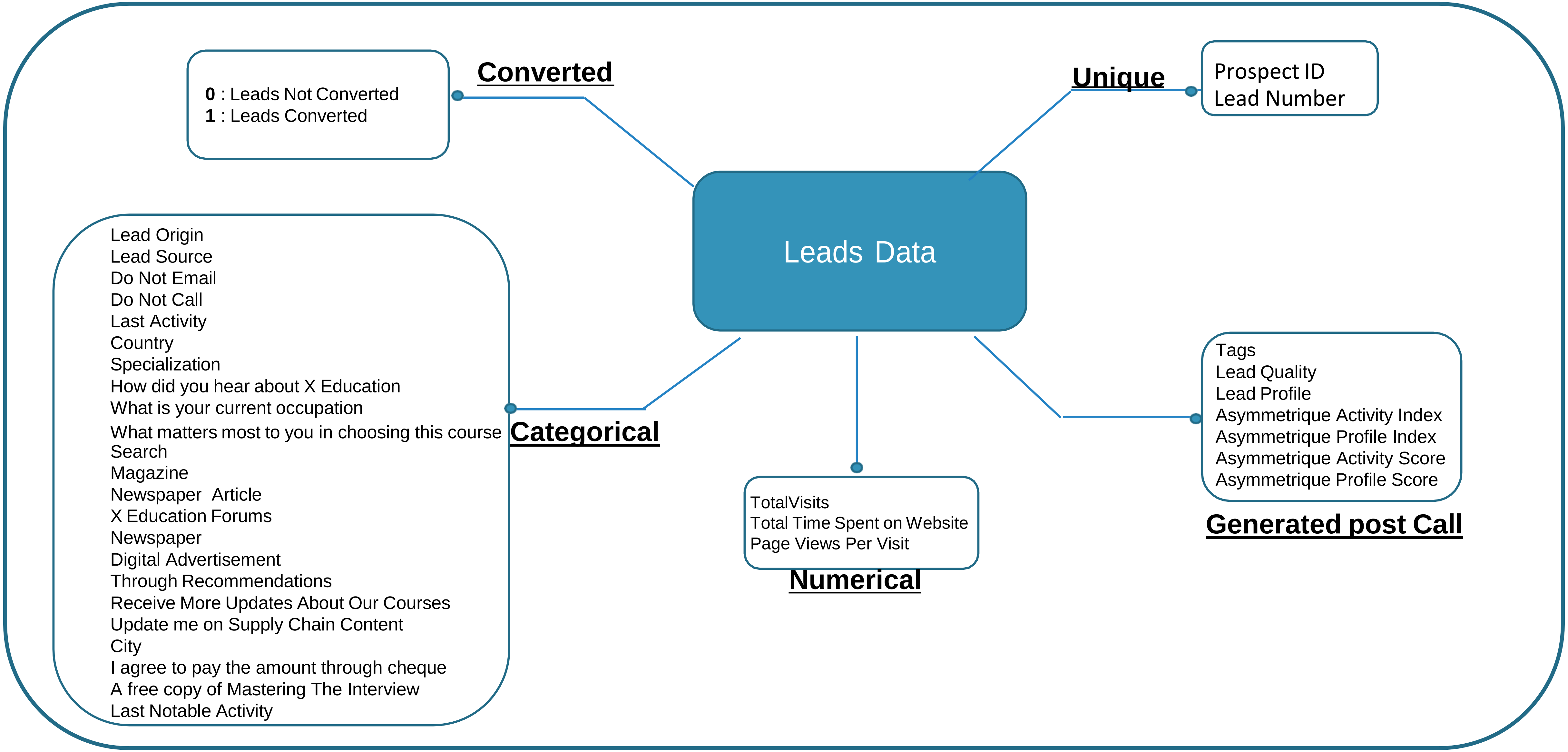
Model Building

- ✓ Use techniques like RFE to perform variable selection.
- ✓ Build a Logistic Regression model with good sensitivity.
- ✓ Check p-value and VIF.
- ✓ Find the optimal probability cutoff.
- ✓ Check the model performance over the test data.
- ✓ Generate the score variable.

Analysis

Conclusion

Data Categorization



Leads Data

Converted

0 : Leads Not Converted
1 : Leads Converted

Unique

Prospect ID
Lead Number

Categorical

Lead Origin
Lead Source
Do Not Email
Do Not Call
Last Activity
Country
Specialization
How did you hear about X Education
What is your current occupation
What matters most to you in choosing this course
Search
Magazine
Newspaper Article
X Education Forums
Newspaper
Digital Advertisement
Through Recommendations
Receive More Updates About Our Courses
Update me on Supply Chain Content
City
I agree to pay the amount through cheque
A free copy of Mastering The Interview
Last Notable Activity

Numerical

TotalVisits
Total Time Spent on Website
Page Views Per Visit

Generated post Call

Tags
Lead Quality
Lead Profile
Asymmetrique Activity Index
Asymmetrique Profile Index
Asymmetrique Activity Score
Asymmetrique Profile Score

Data Cleaning

1. Converting Select values to NaN

For many categorical variables, default value is set as “Select” where the user has not given any input.

Imputing the Value as missing for data standardization

2. Dropping the columns generated after the lead call

Dropping all the columns which were created by the sales team categorising the lead post discussion. These columns are not for much importance for us as they will impact the final lead score

The Columns are:

- Tags
- Lead Quality
- Lead Profile
- Asymmetrique Activity Index
- Asymmetrique Activity Score
- Asymmetrique Profile Index
- Asymmetrique Profile Score

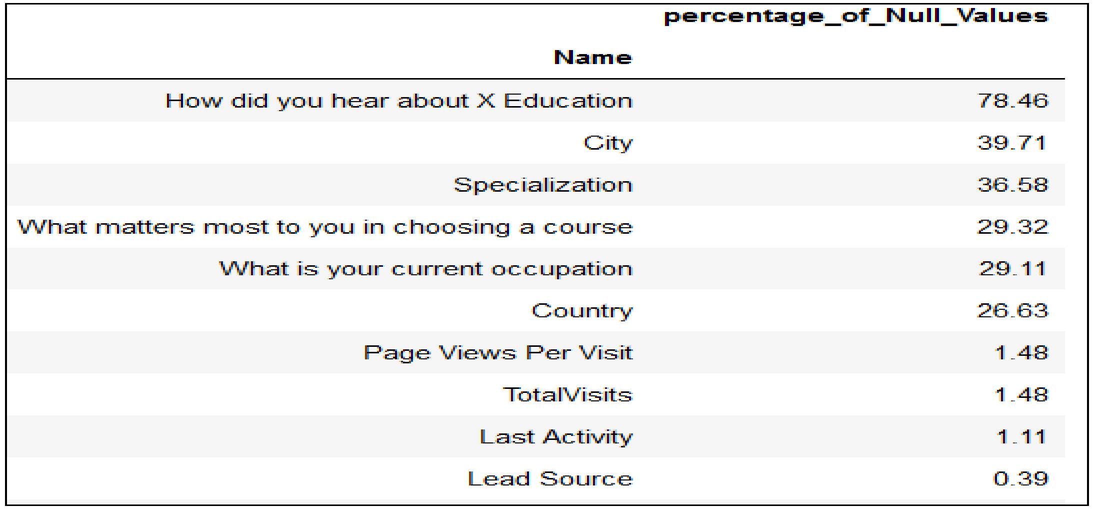
3. Checking for missing values in the dataset:

"How did you hear about X Education" has 78% missing values, Removing the column for further analysis

Below mentioned columns have missing values ranging between 25-40%. After further analysis, we'll decide either to drop the columns or impute the values

- City
- Specialization
- What matters most to you in choosing a course
- What is your current occupation
- Country

For four variables, missing values are below 1.5%



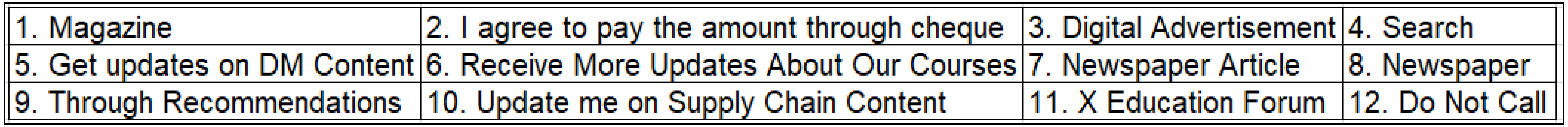
Name	percentage_of_Null_Values
How did you hear about X Education	78.46
City	39.71
Specialization	36.58
What matters most to you in choosing a course	29.32
What is your current occupation	29.11
Country	26.63
Page Views Per Visit	1.48
TotalVisits	1.48
Last Activity	1.11
Lead Source	0.39

Data Cleaning

4. Checking Skewness In Categorical Data

- Highly Skewed Categorical Variables
- ➢ Country : 70% of the rows belong "India" and 29% values are missing. Rest of the countries have value less than 0.8%. Dropping the variable.
- ➢ What matters most to you in choosing a course : 70% of the rows belong to "Better Career Prospects" and approx. 29% of values are missing accounting for 99% of data. Dropping the variable.
- ➢ What is your current occupation : 60% of rows belong to "Unemployed" and 29% of the data is missing. As per business need, keeping the variable
- Binary Variables:

Below mentioned binary variables have 100% or approx. 100% of the data belonging to one category i.e. "NO" making the variable highly skewed. Dropping the below mentioned variables for better analysis.



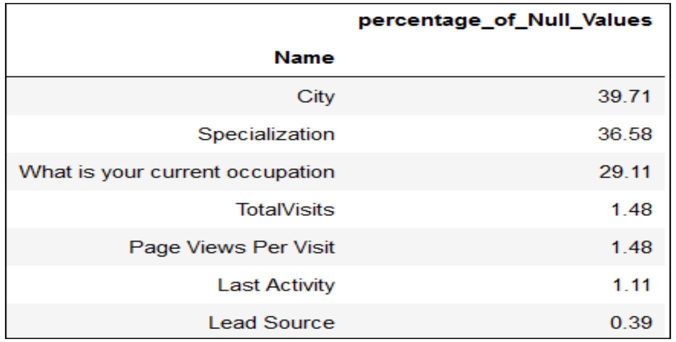
1. Magazine	2. I agree to pay the amount through cheque	3. Digital Advertisement	4. Search
5. Get updates on DM Content	6. Receive More Updates About Our Courses	7. Newspaper Article	8. Newspaper
9. Through Recommendations	10. Update me on Supply Chain Content	11. X Education Forum	12. Do Not Call

5. Combining the Categories

For the Categorical Variables, categories having very little data as compared to other categories, Combing all of them under the "Others" Category for better analysis

6. Checking for missing values and imputing them

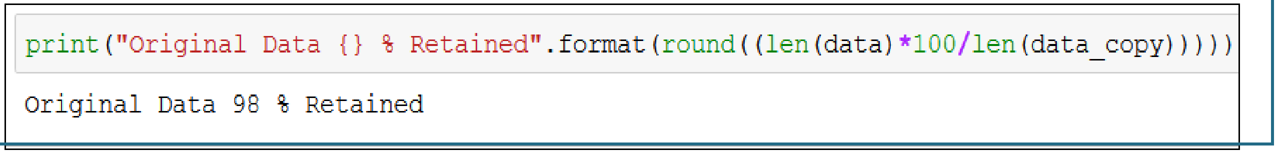
- ➢ "City" variable has 39% missing values. Since this could be an important parameter as per the business understanding, Imputing the missing value with mode i.e. "Mumbai"
- ➢ "Specialization" has 36% missing values. Imputing the data with mode i.e "Umeployed" might impact the dataset, hence, imputing the missing values as "Not_Specified"
- ➢ "What is your current occupation" has 29% missing values. Under occupation, imputing the data with modemight not give accurate results as per business need, Imputing the missing data as "Not_Specified“
- ➢ Rest missing values are under 1.5% so we can drop these rows



Name	percentage_of_Null_Values
City	39.71
Specialization	36.58
What is your current occupation	29.11
TotalVisits	1.48
Page Views Per Visit	1.48
Last Activity	1.11
Lead Source	0.39

7. Calculating the number of rows retained after cleaning the dataset

<u>98% of the rows are retained after cleaning the dataset.</u>



```
print("Original Data {} % Retained".format(round((len(data)*100/len(data_copy)))))
```

```
Original Data 98 % Retained
```

Exploratory Data Analysis

Categorical Data Analysis

1. Lead Origin:

- API and Landing Page Submission have a 30-35% conversion rate but a count of leadoriginated from them are considerable.
- Lead Add Form has more than 90% conversion rate but the count of lead are not very high.
- Lead Import is very less in count.

2. Lead Source:

- Google and Direct traffic generates maximum number of leads.
- Conversion Rate of "Others" is high.

3. City:

- Most leads are from Mumbai with around 30% conversion rate.

4. Last Activity:

- Most of the lead have their Email opened as their last activity.
- Conversion rate for leads with last activity as SMS Sent is almost(approx. 1700)that is 60%.

5. Last Notable Activity:

- Count for Modified is the highest.
- Conversion Rate for SMS sent is high.

6. What is your Current Occupation:

- Count of unemployed leads is high with more than 70% conversion rate.
- Working professionals have great conversion rate.

7. Specialization:

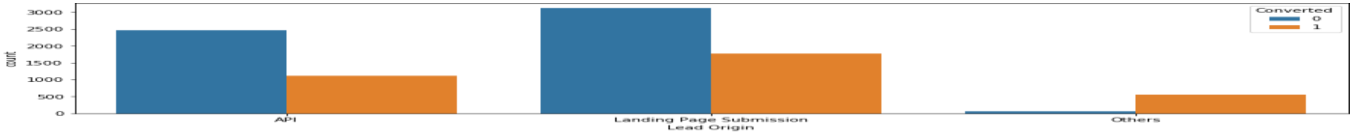
- None of the specialization has a major impact
- The highest count are for Others(rest of the categories under Specialization which had very less count) and Not-specified

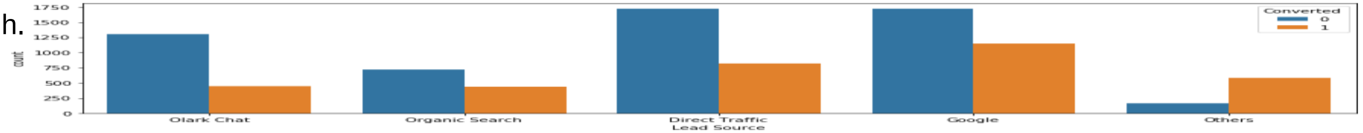
8.A Free Copy of Mastering the Interview:

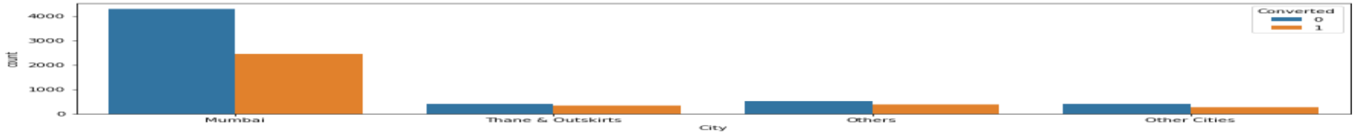
- A free copy attracts few customers.

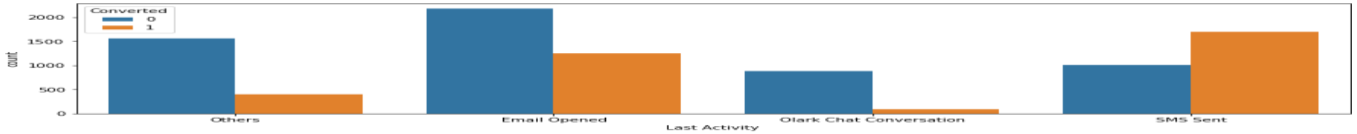
9. Do Not Email:

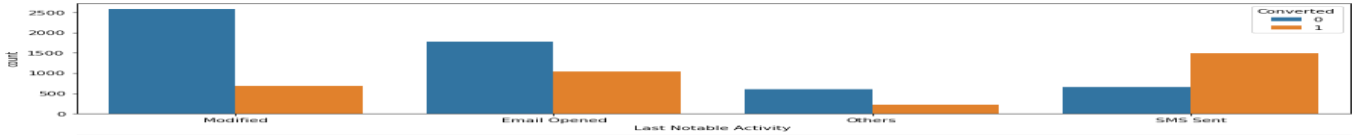
- Customers do not prefer to be called or emailed.
- Very few customers look forward to an email follow- up.

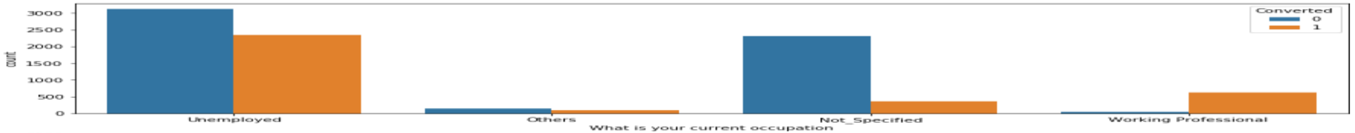


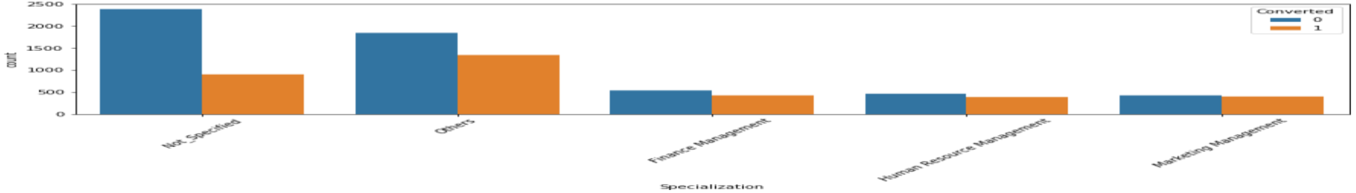


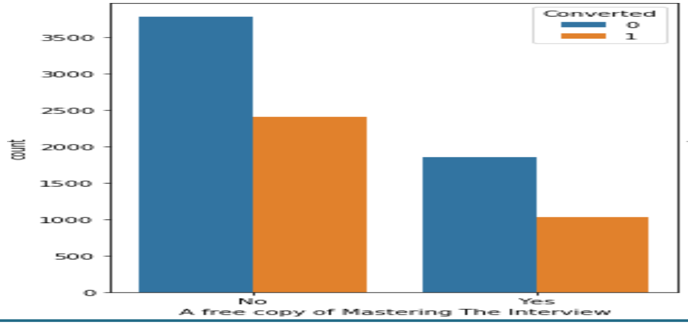


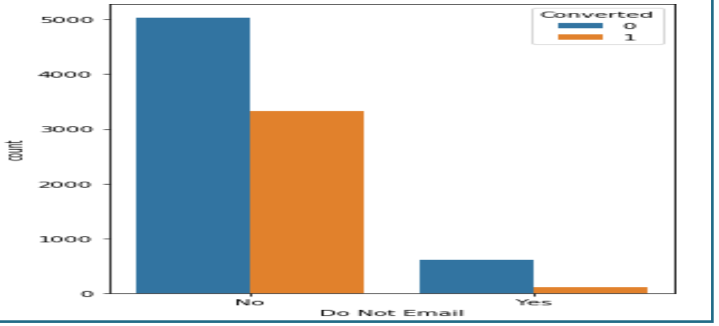






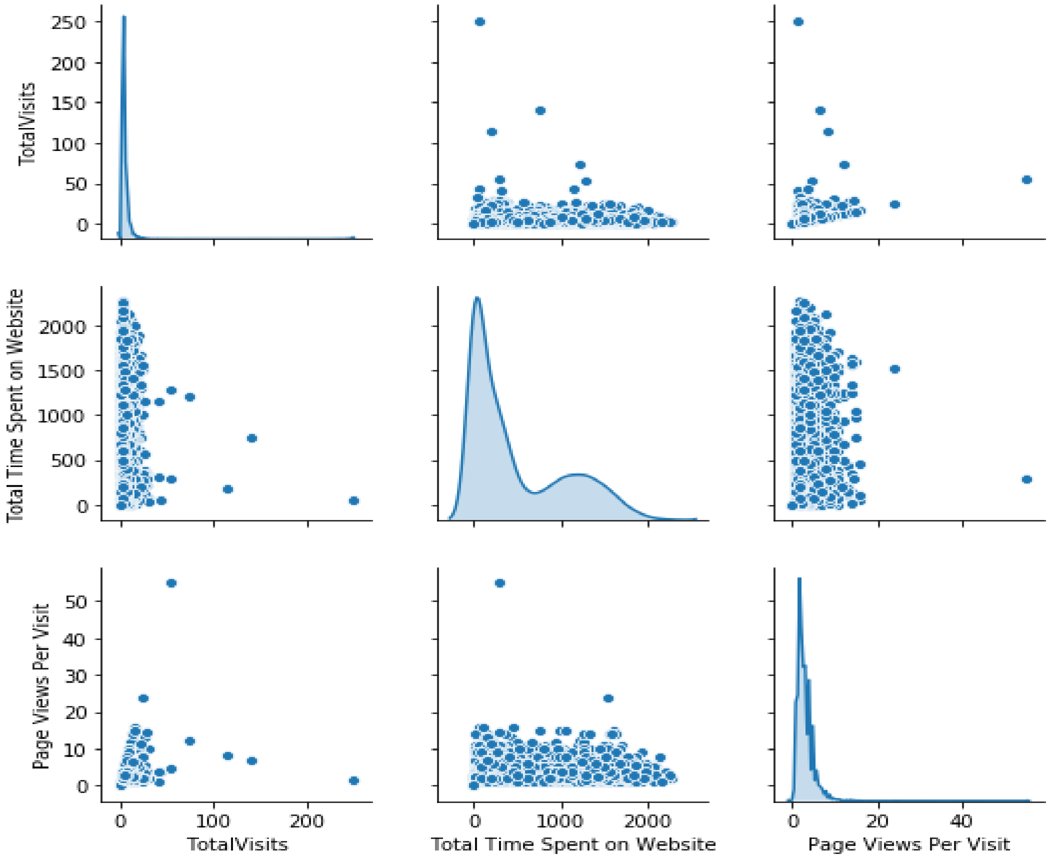






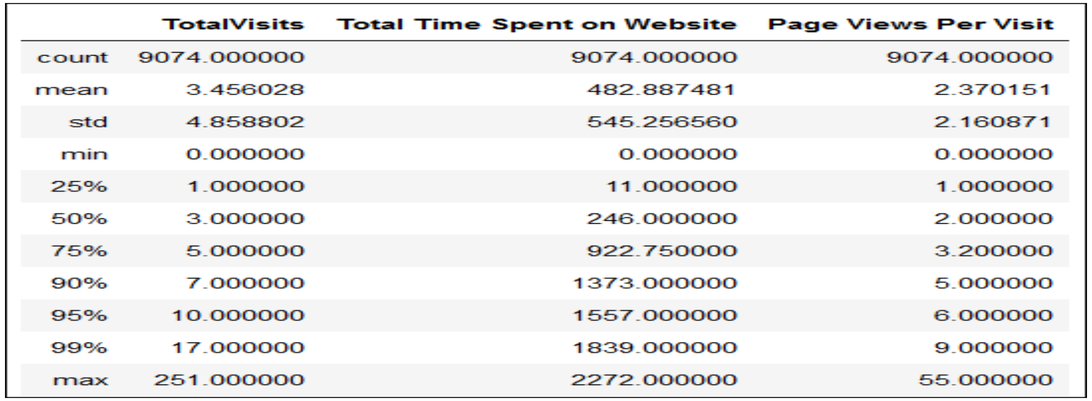
Exploratory Data Analysis and Outlier Treatment

Numerical Variables

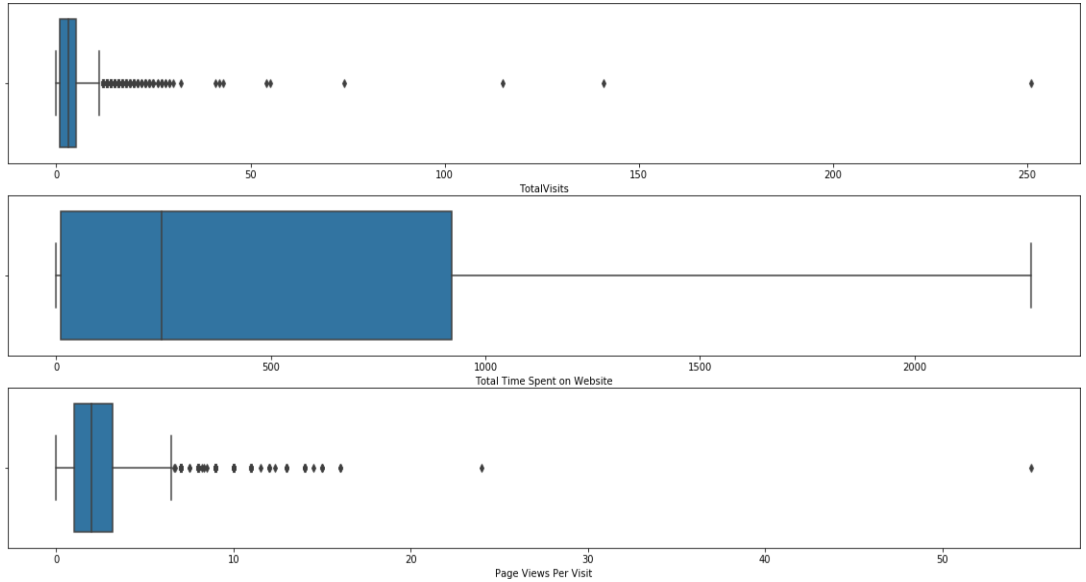


- Total Visits and Page Views Per Visit are linearly related
- Total Time Spent on website and Page Views per visit are related but clustered

Outlier Analysis and Treament



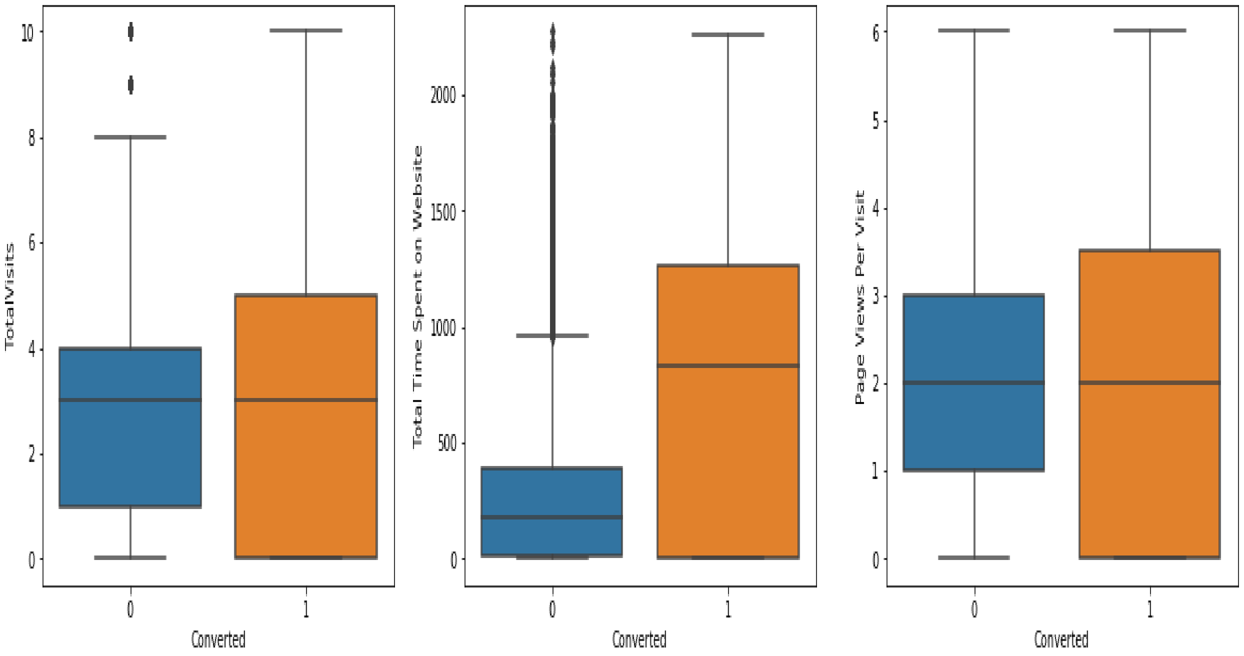
	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	9074.000000	9074.000000	9074.000000
mean	3.456028	482.887481	2.370151
std	4.858802	545.256560	2.160871
min	0.000000	0.000000	0.000000
25%	1.000000	11.000000	1.000000
50%	3.000000	246.000000	2.000000
75%	5.000000	922.750000	3.200000
90%	7.000000	1373.000000	5.000000
95%	10.000000	1557.000000	6.000000
99%	17.000000	1839.000000	9.000000
max	251.000000	2272.000000	55.000000



- Total Visits have a lot of outliers at upper end
- There are no outliers in Total Time Spent on Website
- Page Views per Visit has outliers in the upper end
- Performing mid-range capping of both the outliers

Exploratory Data Analysis

Boxplot for numerical variables:



1. Total Visits:

- Median for converted and not converted leads are the same. Nothng conclusive can be said on the basis of plot.

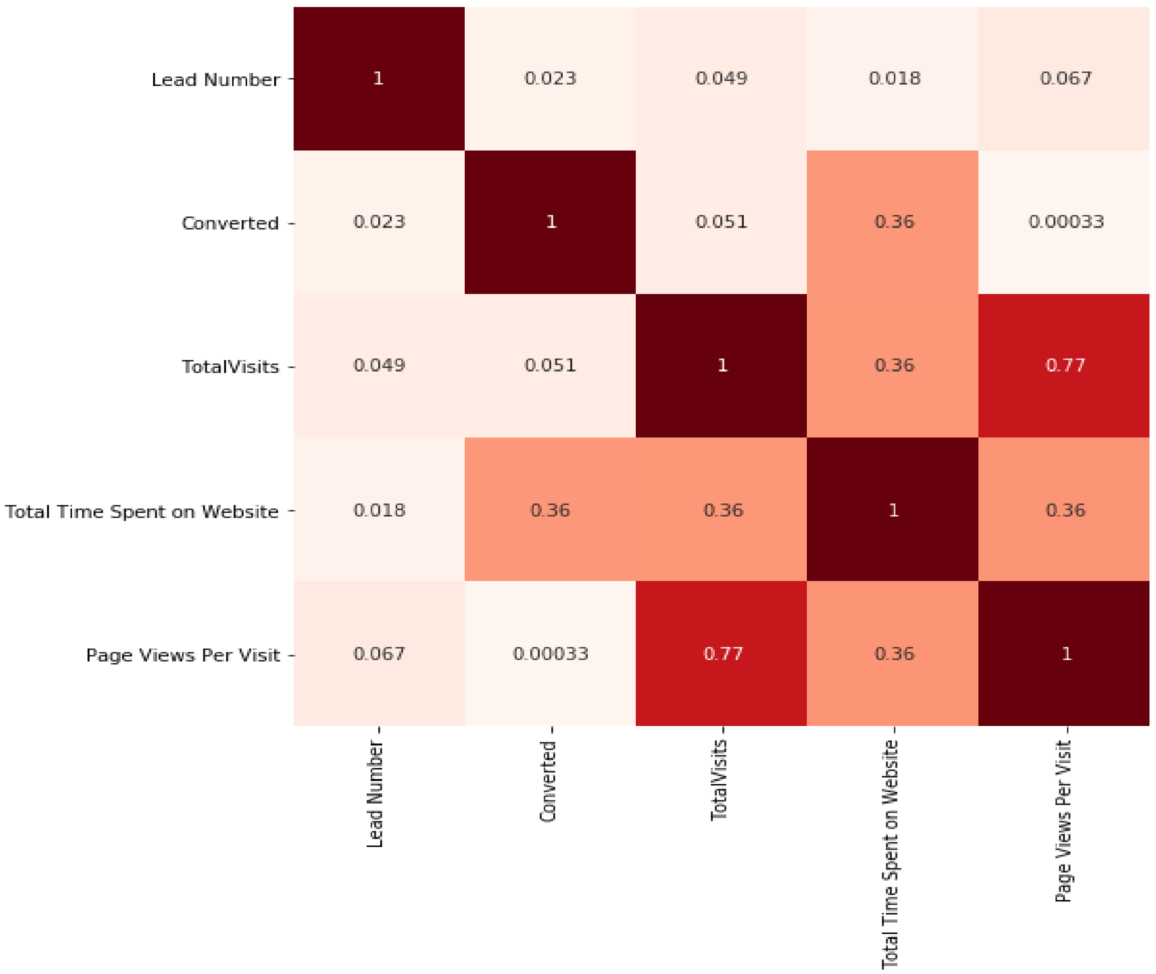
2. Total Time Spent on Website:

- Leads spend more time on the websites are more likely to be converted. Website should be made more attractive and engaging to make leads spend more time.

3. Page Views Per Visit:

- Median for converted and not converted leads is the same, not indicating anything conclusive on the basis of plot.

Correlation Matrix:



1. A significant correlation (0.36) with converted can be seen with Total time visit on the Website.
2. A weak correlation(0.00033) is seen between converted and Page views per Visit

Data Preparation

1. Converting Binary variables:

For both the binary variables, i.e., 'Do Not Email’and 'A free copy of Mastering The Interview’, yes were converted to 1 and no were converted to 0

2. Creating dummy for all the categorical variable:

Below steps were followed

- Create Dummy variable
- Drop original variable for which the dummy was created
- Drop first dummy variable for each set of dummies created.

3. Diving dataset into test and train:

Importing the library and diving the dataset such that 70% is allocated to train set and 30% is for test set

Choosing the random_state=69 as this will always perform the same split whenever the command is executed

4. Feature Scaling

We used the Standard Scaling to scale the original numerical Variables.

Standard-Scaler scales the features around the centre with mean 0 and with a standard deviation of 1.

Model Building

1. Using the Recursive Feature Elimination, we went ahead and selected the 20 top important features.
2. Using the statistics model generated, we recursively tried looking at the P-values and VIFs in order to select the most significant features and dropped the insignificant ones.
3. Due to low Multicollinearity between the predictors and significant p-values, considering the model 7 as our final model.
4. Finally, we arrived at the 15 most significant variables. The VIF’s for these variables were also found to be good (all less than 5).

Model Building

Predicting the values on train dataset

We then created the dataframe with the actual Converted values and the predicted probabilities taking cutoff as 0.5. i.e. if the probability is greater than 0.5, the predicted value will be 1 else 0.

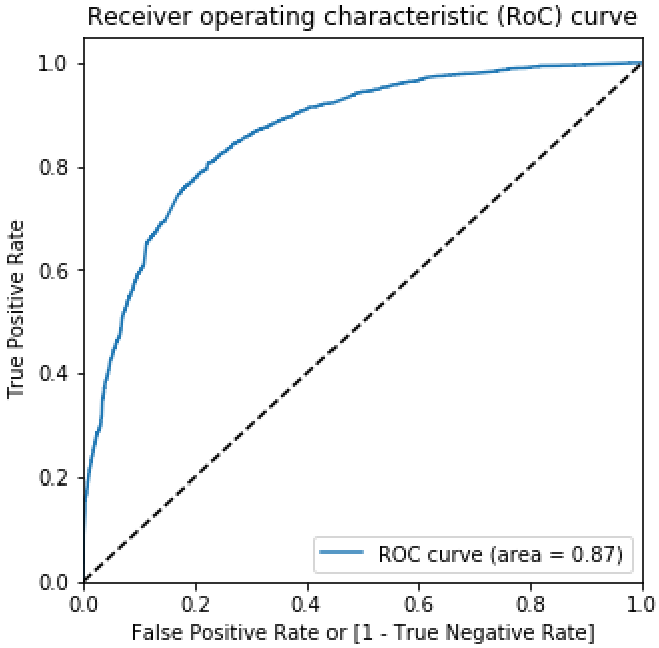
Based on this, we got the results as:

- **Accuracy : 79.64%**
- **Sensitivity: 66.47%**
- **Specificity : 87.79%**

The model was not performing as ideal and we need to have an optimal cut-off which would enhance the model’s accuracy and other metrics

Plotting ROC Curve:

ROC curves is used for determining the best cut-off value for predicting whether a new observation is "failure" (Not_converted) or a "success" (Converted)

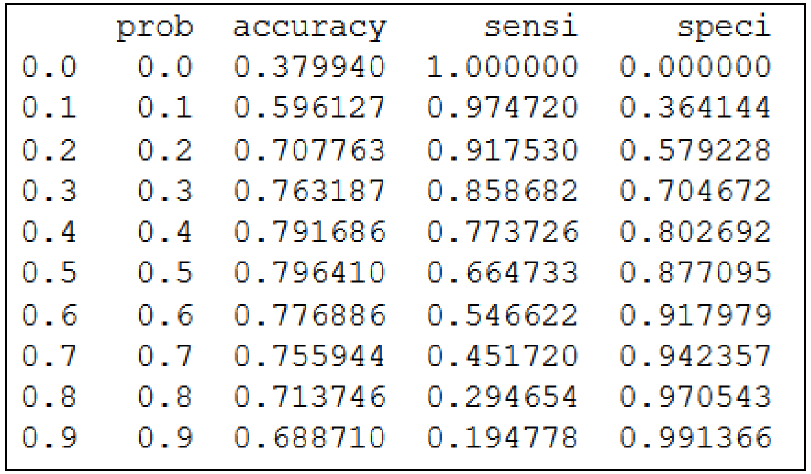


The ROC Curve should be a value close to 1. We are getting a good value of 0.87 indicating a good predictive model.

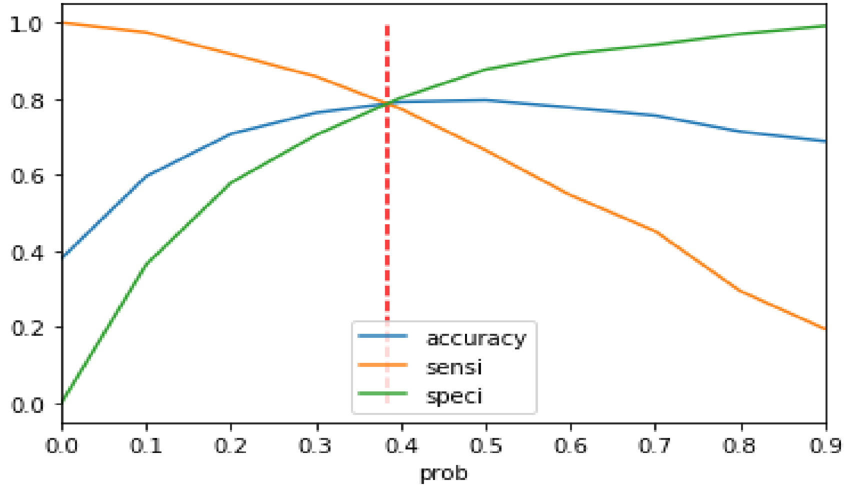
Model Building

Finding the Optimal Cutoff Point

1. Then we plotted the probability graph for the ‘Accuracy’, ‘Sensitivity’, and ‘Specificity’ for different probability values.
2. The intersecting point of the graphs was considered as the optimal probability cutoff point. The cutoff point was found out to be **0.38**



```
     prob  accuracy     sensi     speci
0.0   0.0  0.379940  1.000000  0.000000
0.1   0.1  0.596127  0.974720  0.364144
0.2   0.2  0.707763  0.917530  0.579228
0.3   0.3  0.763187  0.858682  0.704672
0.4   0.4  0.791686  0.773726  0.802692
0.5   0.5  0.796410  0.664733  0.877095
0.6   0.6  0.776886  0.546622  0.917979
0.7   0.7  0.755944  0.451720  0.942357
0.8   0.8  0.713746  0.294654  0.970543
0.9   0.9  0.688710  0.194778  0.991366
```



Assessing the model based on the new cut-off obtained and providing a lead score to each lead

After optimising the performance on train data set, we can see above the model seems to be performing well. The ROC curve has a value of 0.87, which is very good. We have the following values for the Train Data:

- **Accuracy : 78.85%**
- **Sensitivity : 78.94%**
- **Specificity : 78.80%**

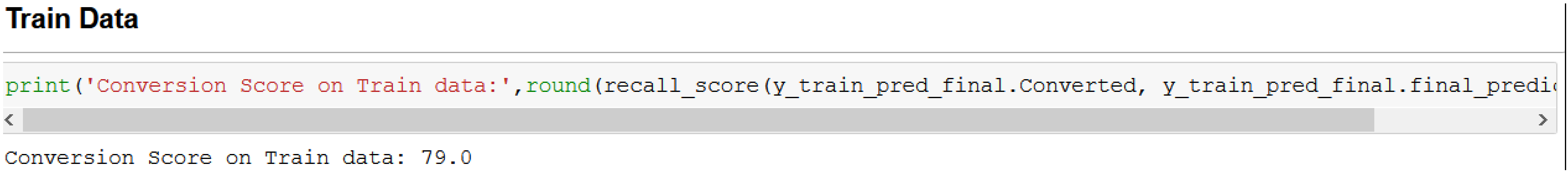
Making prediction in the test data set:

Implementing the learnings on the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found as:

- **Accuracy : 78.41%**
- **Sensitivity : 77.10%**
- **Specificity : 79.18%**

Model Performance

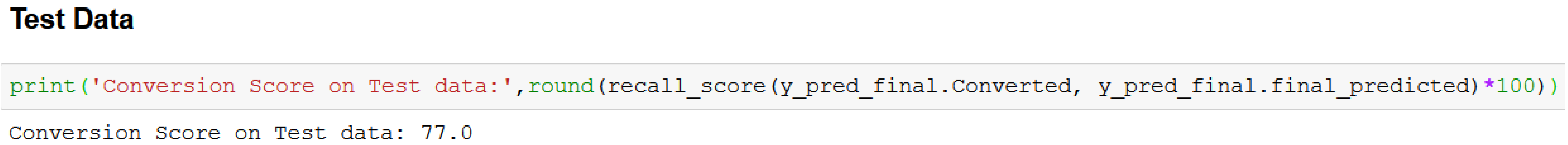
Generating the final Probabality (Lead) Score



Train Data

```
print('Conversion Score on Train data:',round(recall_score(y_train_pred_final.Converted, y_train_pred_final.final_predi
```

```
Conversion Score on Train data: 79.0
```



Test Data

```
print('Conversion Score on Test data:',round(recall_score(y_pred_final.Converted, y_pred_final.final_predicted)*100))
```

```
Conversion Score on Test data: 77.0
```

Based on the model build, we are getting the final metrics are mentioned below:

➢ Train Data Set :

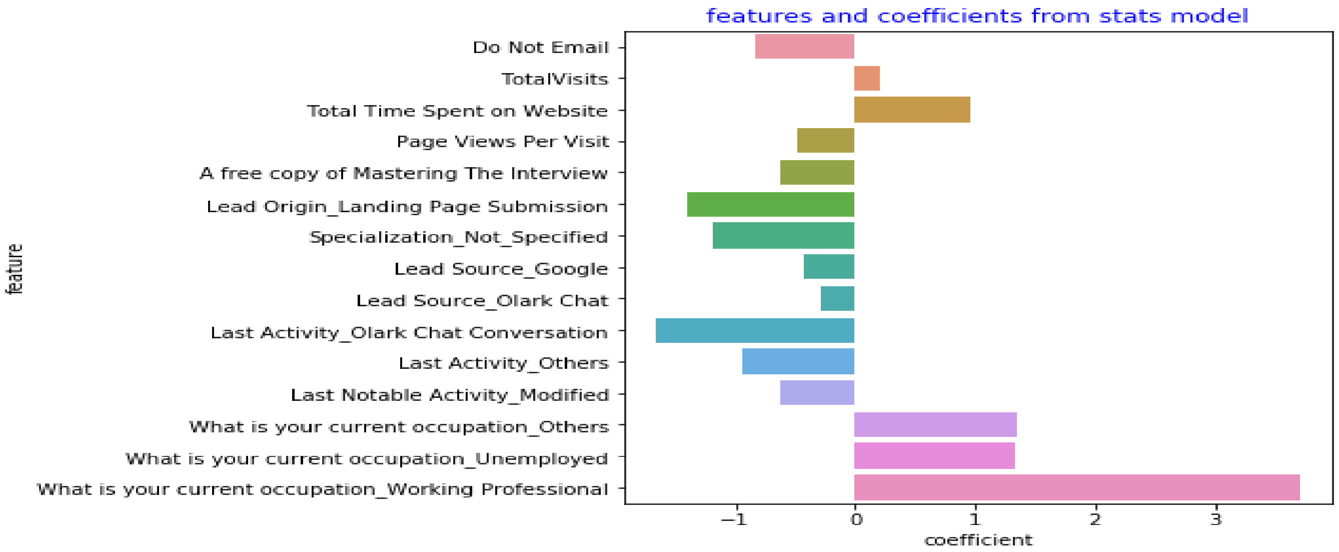
- **Accuray: 78.85%**
- **Sensitivity: 78.94%**
- **Specificity: 78.80%**
- **Score: 79%**

➢ Test Data Set :

- **Accuray: 78.41%**
- **Sensitivity: 77.10%**
- **Specificity: 79.18%**
- **Score: 77%**

The Model predicts the Conversion Rate well and we will give our recommendations based on this model

The top Features of the model along with their Coefficients



Conclusion and Recommendations

From the model, we have optimized the conversion rate of approx. **38%** to a final conversion rate of **79%.**

The conversion went up by eliminating the features which were having a lot of noise in data and were not helping for lead conversion.

Based on this model, the sales team will know the key indicators and the ideal lead score which they should analyze before reaching out to the customer.
Along with boosting the conversion rate, this model will also help in increasing efficiency of the team and provide them with a more focused approach of lead conversion.

The top 3 variables which contribute most towards the probability of a lead getting converted are:

- What is your current occupation
- Total Time Spent on Website
- TotalVisits

Recommendations

- ➢ Total time spent on the website and the total no. of visits is an important factor for lead conversion, the company should invest on making their website more attractive and engaging in order to enhance the conversion rate.
- ➢ The company should focus on providing attractive offers to the currently unemployed people as these people have the maximum lead generation count. Also, working professional have the maximum conversion rate which means they are also beneficial for the company.
- ➢ Almost 100% of the leads have responded negatively as whether they have seen the companies advertisement on various platform. The marketing team should focus more on this as current spending on advertisement is not helping in lead conversion.
- ➢ The company should focus more on lead sourcing through Google and Olark Chat system.