Lead Score Case Study

By Utkarsh Raj

Problem Statement

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.

Understanding the data

Provided data is Leads.csv

a. It contains 1940 rows and 37 columns in which 7 columns are numeric and 30 are categorical.

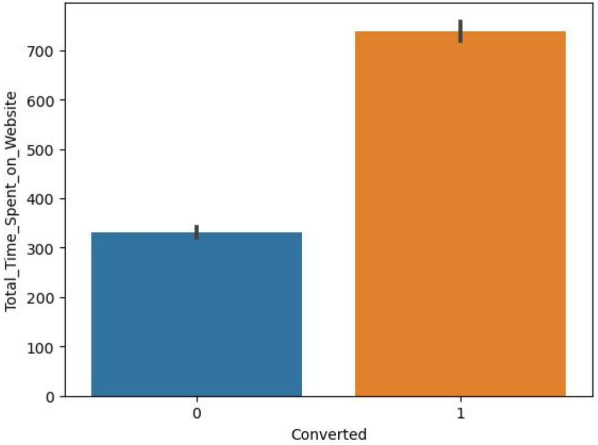
b. Conversion rate of the provided data is 39%.

c. Leads Data Dictionary.csv is also provided which is the metadata of the main dataset.

Data Cleaning and Preparation

1. We replace ‘select’ with nan since select is default value for some columns which is equivalent to null values.
2. Drop all the unnecessary columns
    a. Asymmetrique_Profile_Index
    b. Asymmetrique_Activity_Index
    c. Asymmetrique_Activity_Score
    d. Asymmetrique_Profile_Score
    e. Lead_Profile
    f. Tags
    g. Lead_Quality
    h. How_did_you_hear_about_X_Education
    i. City
    j. Lead_Number
3. Now, we check the unique values in the columns and drop the columns containing just 1 unique value.

Check for good variance between Target and other Continuous Variables



Converted VS Total Time Spent on Website can be a good predictor.

More Data Cleaning

We replace words like ‘google’ with ‘Google’ since this created occurrence of 2 google in thee dataset.

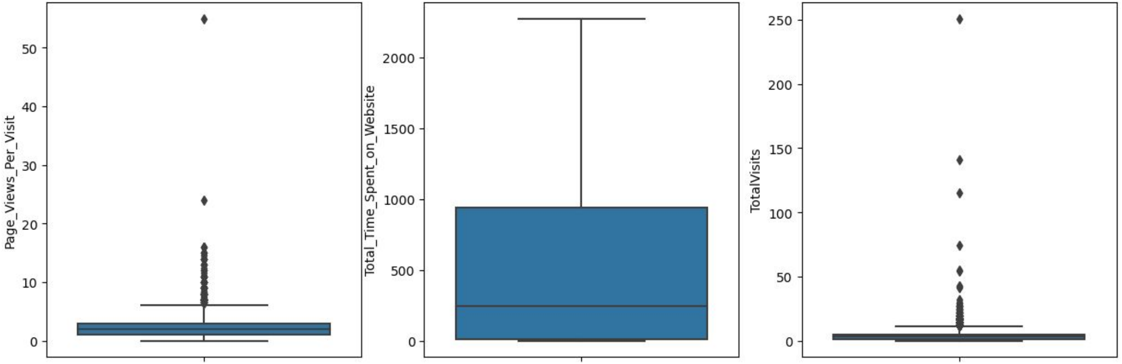
Also, replace the insignificant values with ‘others’ to group them together.

Replace all the remaining nulls from important variables with not_declared to fill the nulls.

Replace null values from country with not_declared and other than india with ‘Outside India’.

Finally, drop the remaining insignificant variables.

Univariate Analysis



Check for outliers from the potential important variables, drop remaining nulls and normalise the numerical values.

Bivariate Analysis: Checking Correlation

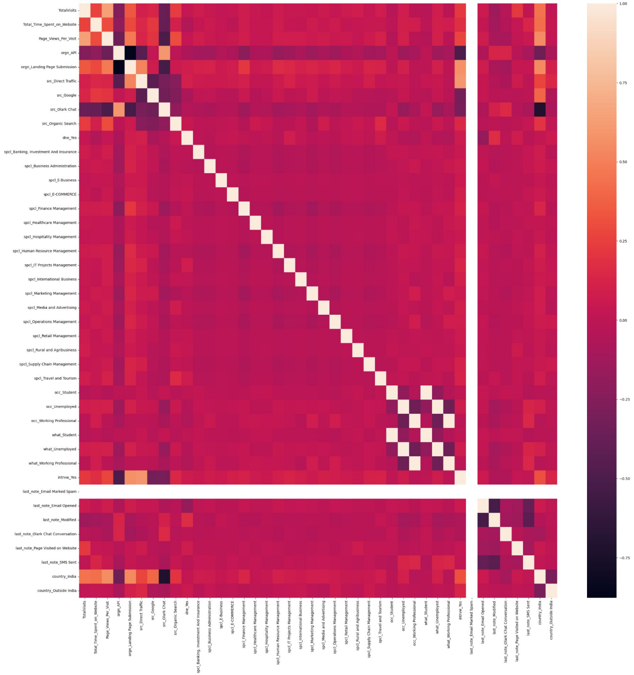
Columns with positively highly correlated with

each other:

1. Total Visits

2. Total Time Spent on Website

3. Page Views Per Visit



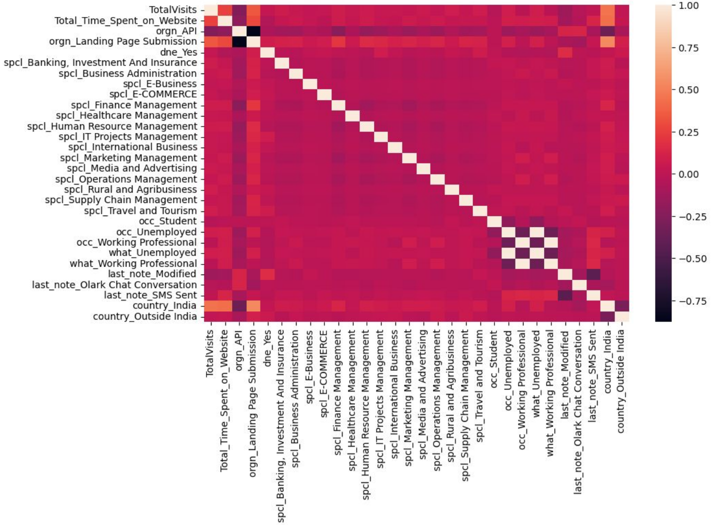
Heatmap for the chosen columns

Choose columns for the training data

using RFE (Recursive Feature Elimination)

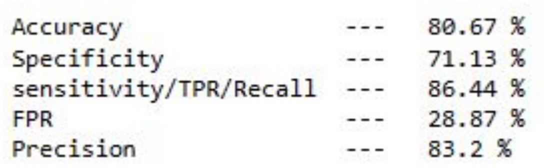
Method.

Checked correlation using the heatmap.



Accuracy

Model gives the accuracy of approx 81% with the testing data.



```
Accuracy                  ---   80.67 %
Specificity               ---   71.13 %
sensitivity/TPR/Recall    ---   86.44 %
FPR                       ---   28.87 %
Precision                 ---   83.2 %
```

Conclusion

This model gives an accuracy of over 81% in identifying the hot leads for the client with lesser time utilised with considerably less resources.

Using the model will help in

a. Shorter sales cycle allowing the client to focus on other parts of the business.
b. Increased marketing effectiveness.
c. Help in sales forecasting allowing the client to plan accordingly.