Lead Scoring Case study

Akshita Khare

Naveli Nandan

There are quite a few goals for this case study.

- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use

Goals of the Case Study

Data cleaning and data manipulation.

1. Check and handle duplicate data.
2. Check and handle NA values and missing values.
3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
4. Imputation of the values, if necessary.
5. Check and handle outliers in data.

EDA

1. Univariate data analysis: value count, distribution of variable etc.
2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

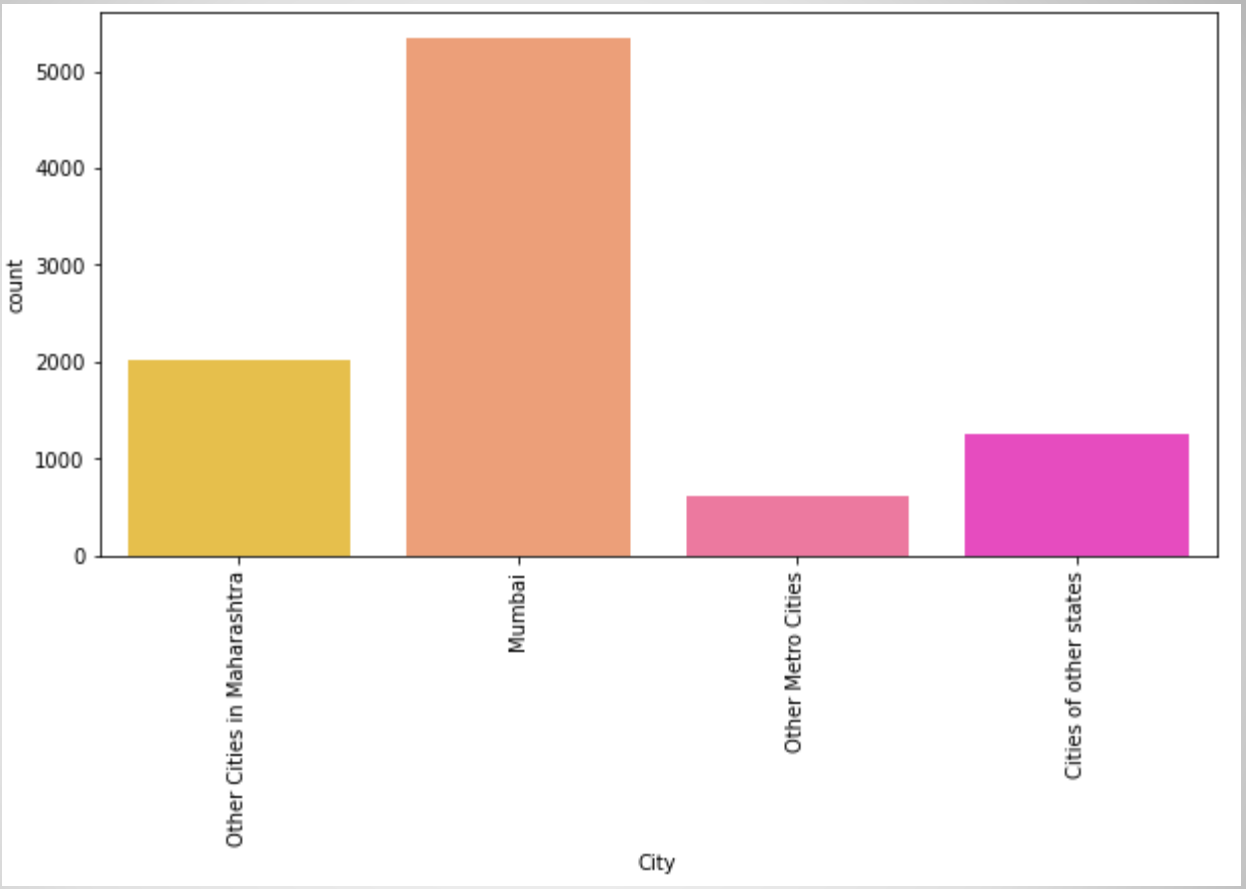
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations

Solution Methodology

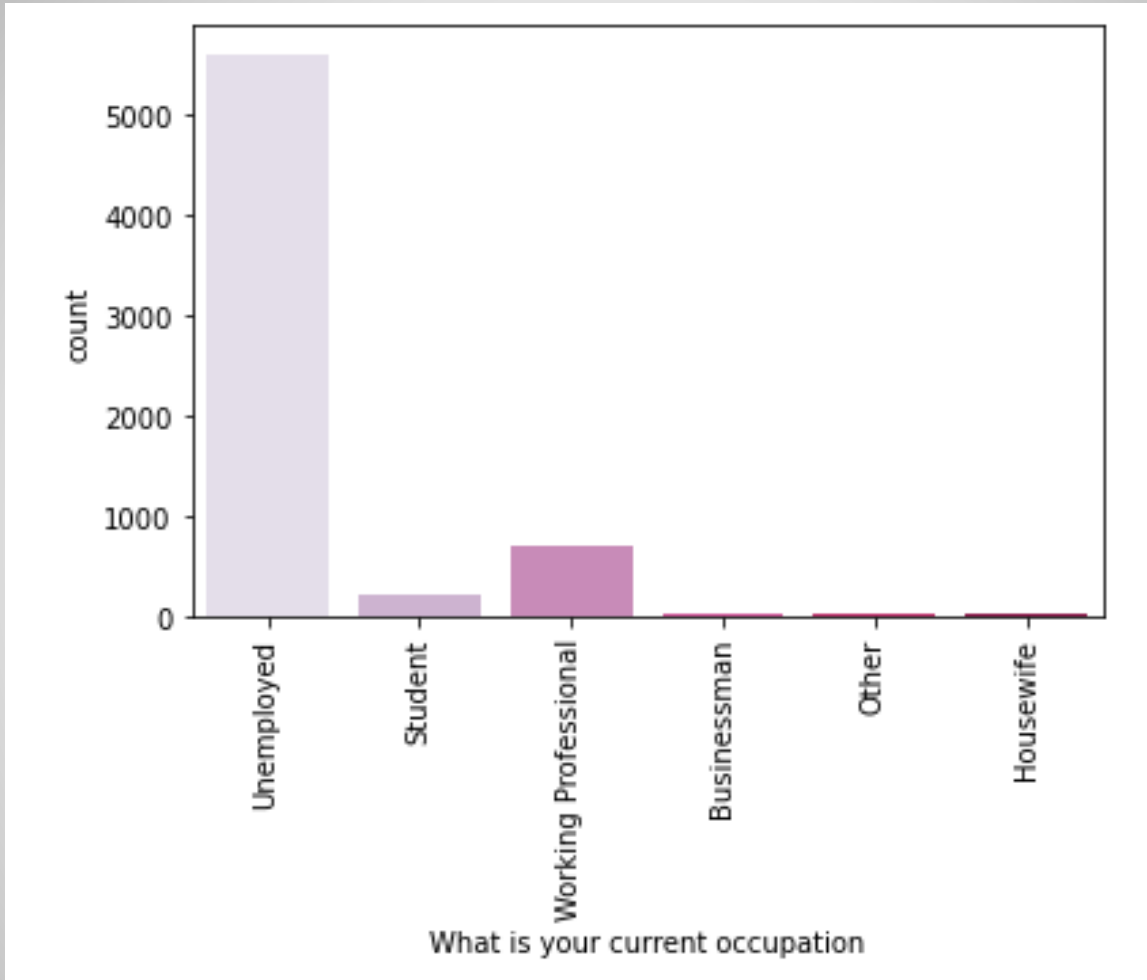
Data Manipulation

- No of Rows – 9240 and No of columns – 37
- Statistical aspects of the dataframe
- Checking for missing values (column-wise)
- Tags , Last Activity and Last Notable Activity contain remarks from the sales executive who make calls to the potential customers. We can't use this data generated by sales team for the purpose of building a model that calculates lead score. So these columns can be dropped.
- Checking the percentage of missing values
- Dropping the columns with 40% or more missing values
- Imputing missing values in columns with high number of null values¶
- The distribution of the data is heavily skewed. Therefore we can drop such columns.

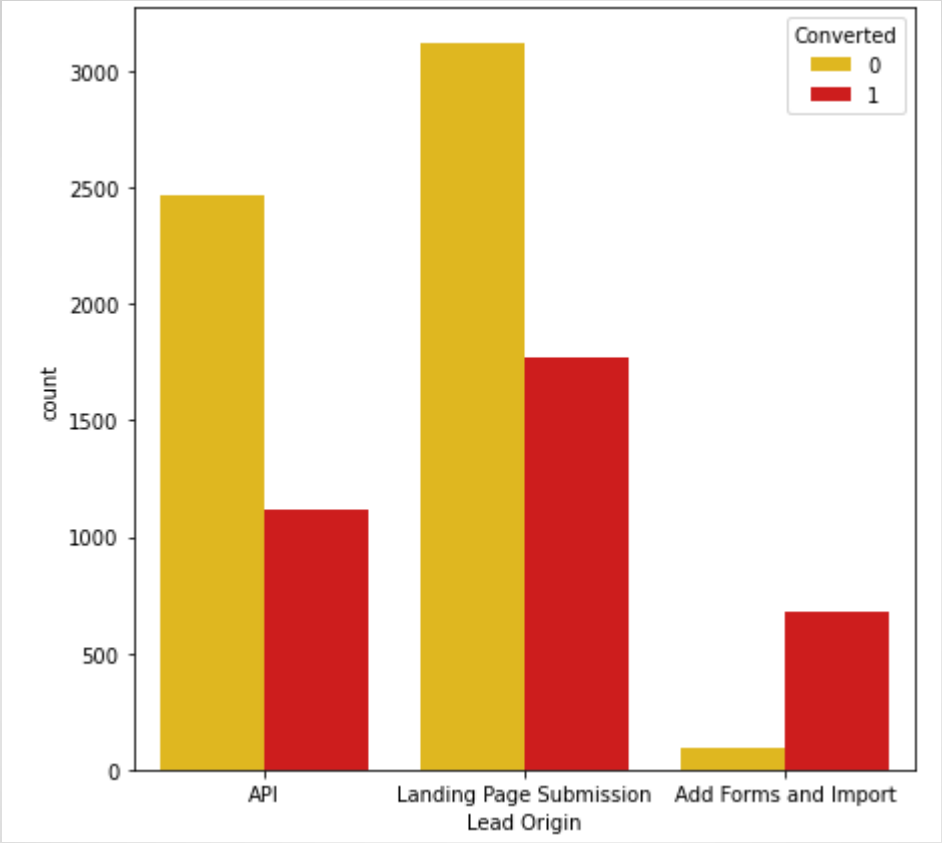
EDA



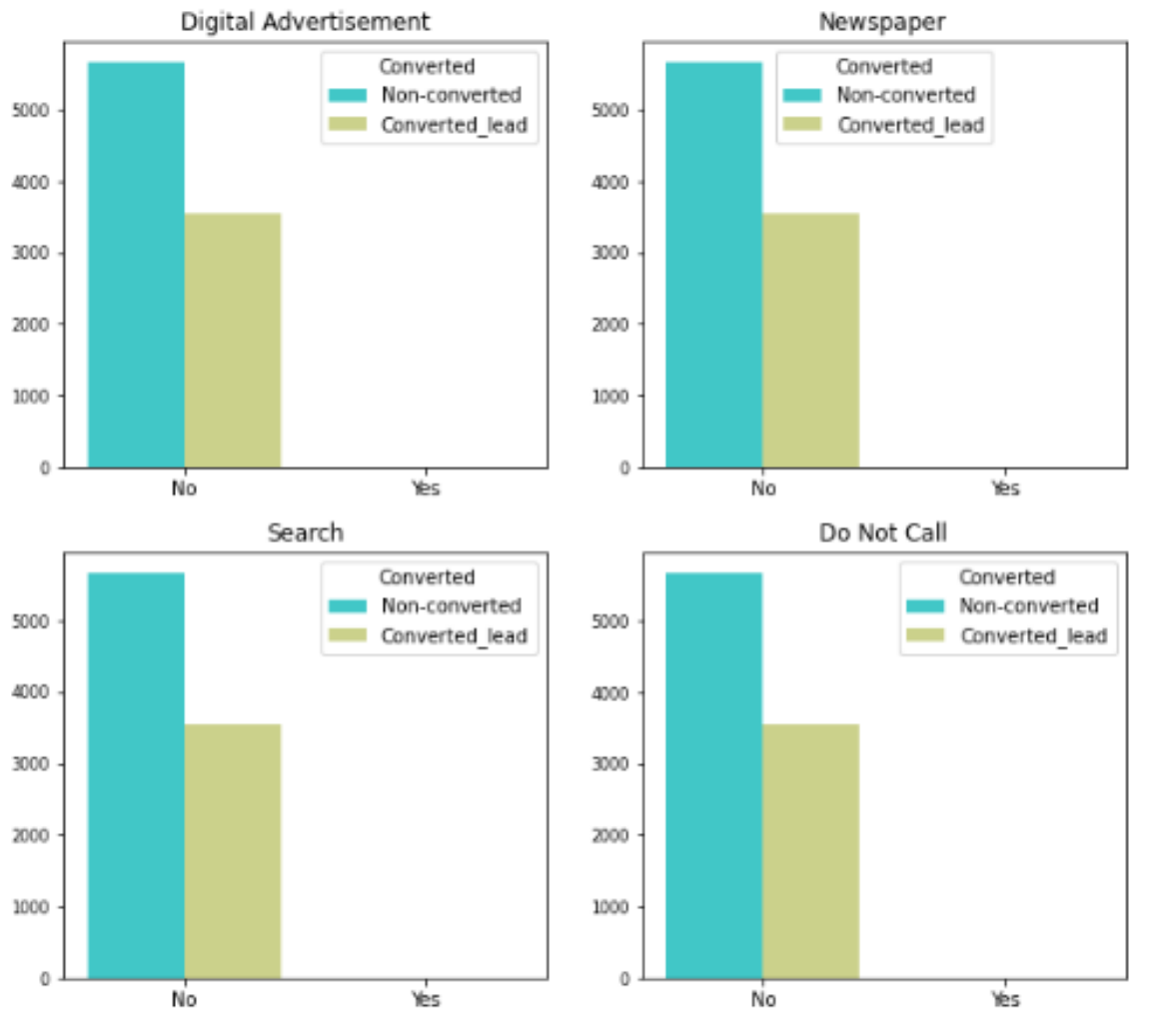
Mumbai have more leads than other cities

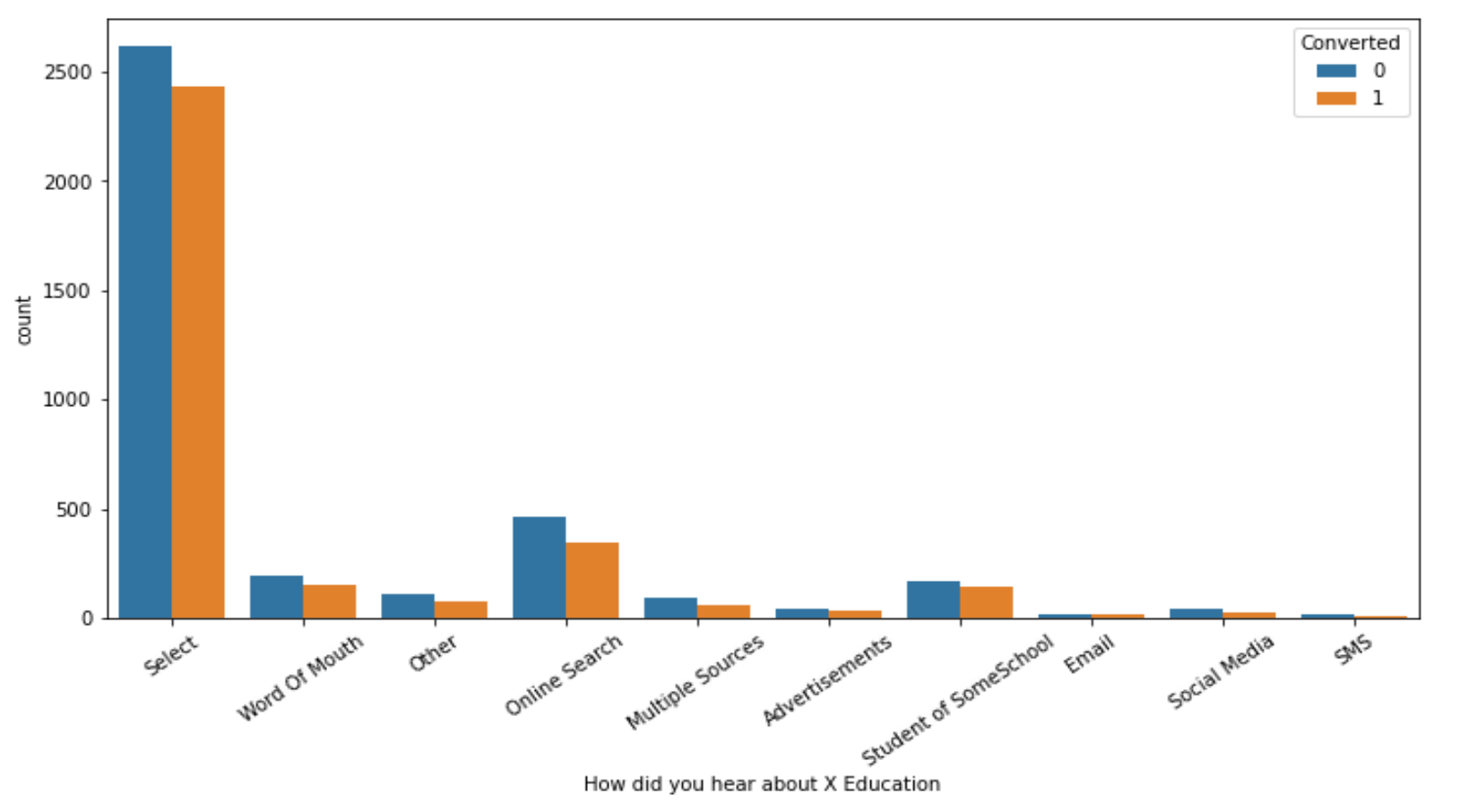


Unemployed people are more than others



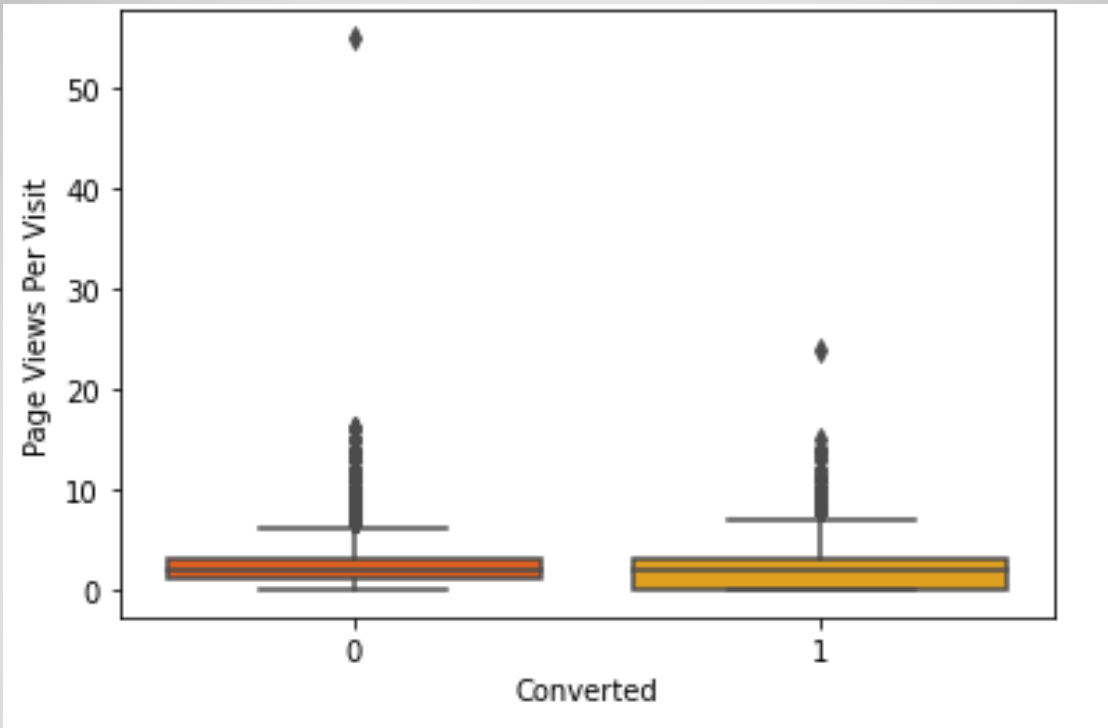
Categorical Variable Relation EDA


Digital Advertisement
Converted
Non-converted
Converted_lead
5000
4000
3000
2000
1000
0
No
Yes
Newspaper
Converted
Non-converted
Converted_lead
5000
4000
3000
2000
1000
0
No
Yes
Search
Converted
Non-converted
Converted_lead
5000
4000
3000
2000
1000
0
No
Yes
Do Not Call
Converted
Non-converted
Converted_lead
5000
4000
3000
2000
1000
0
No
Yes



- As we can see the number of values for not selected are quite high (nearly 98% of the Data), this column can be dropped

Imbalance data

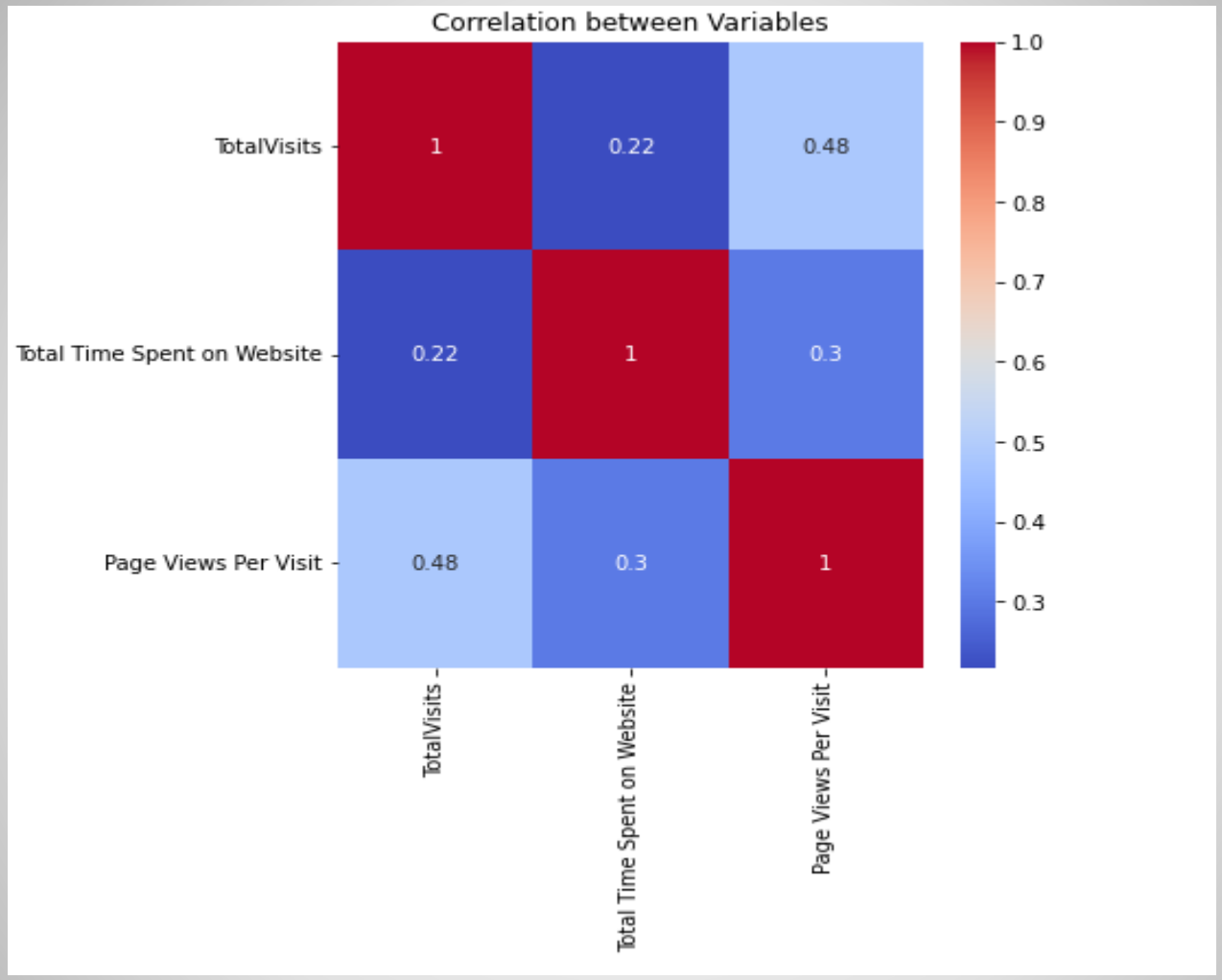


The trend here is similar to what we saw for the TotalVisits column. Outlier treatment will be done for all the columns in the data preparation section.

Outlier Treatment

Data Conversion

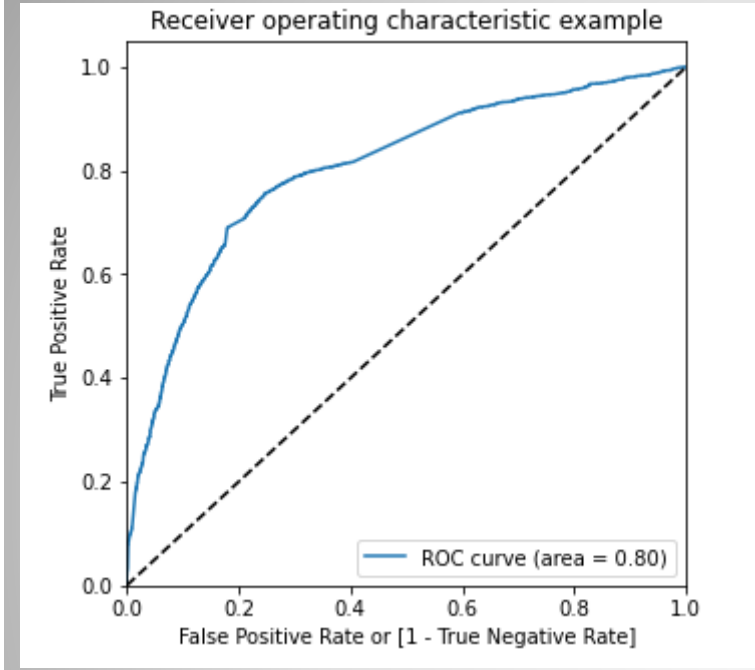
- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Looking at correlations
- Dropping highly correlated dummy variables
- Check for the total Rows for Analysis
- Check for the total Columns for Analysis.

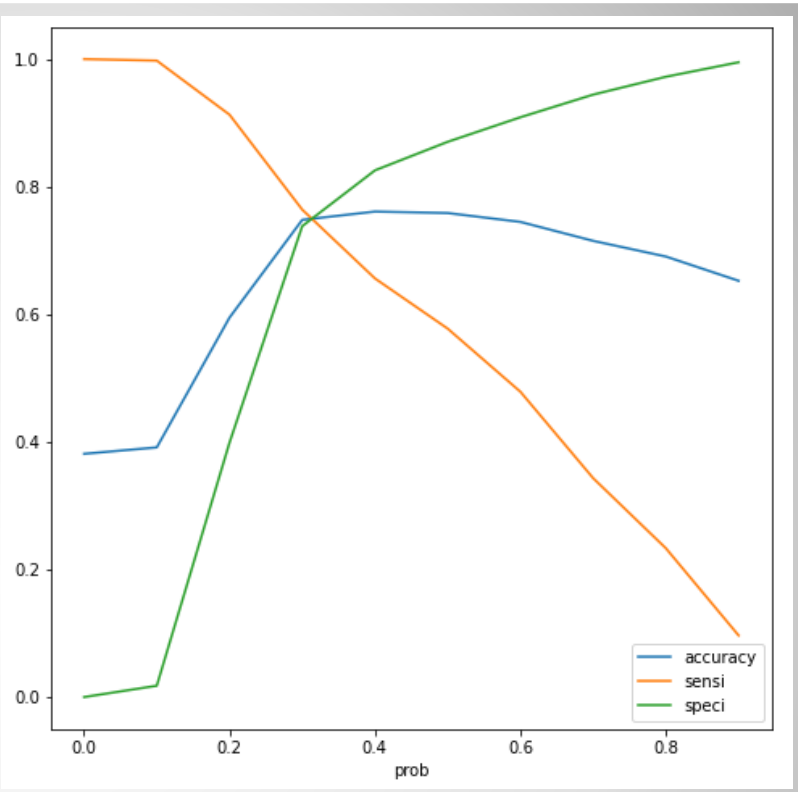


correlations

Model Building

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 10 variables as output
- Building Model by removing the variable whose p-value is greater than 0.05 and vif value is greater than 5
- Predictions on test data set
- Overall accuracy 74.38%





ROC Curve

- The ROC Curve should be a value close to 1. We are getting a good value of 0.80 indicating a good predictive model.
- From the curve above, 0.3 is the optimum point to take it as a cutoff probability.

The variables that mattered the most are:

- Total time spend on Website.
- Total visits.
- Page views per visit

- The lead source was:

1. Google
2. Olark chat
3. Organic search
4. Reference

Conclusion

Thank You