LEAD SCORING CASE STUDY

PROBLEM STATEMENT

- ➢ An education company named X Education sells online courses to industry professionals.
- ➢ They get a lot of leads through various sources, but their lead conversion rate is very poor, for e.g., in 100 only 30 leads get converted
- ➢ To make this process more efficient, the company wishes to identify the most potential leads
- ➢ If they successfully identify this set of potential leads their sales team will only focus on these instead of wasting time in calling everyone

BUSINESS OBJECTIVE

- ➢ X education wants to know most Promising leads
- ➢ They want to build the model that will identify the Promising leads
- ➢ And they want to deploy the model for future use

SOLUTION METHODOLOGY

DATA PREP.

Check and Handle Duplicate data, Check and Handle Missing data, Drop Unnecessary Columns, Imputation of values

DATA VISUALIZATION AND EDA

Check and Handle Outlier

Univariate Data analysis

Bivariate Data

analysis

DATA CONVERSION

Feature Scaling, dummy variables and encoding of data

Test and train split

MODEL BUILDING

Building the Logistic Model

Drop the columns

according to p- value and rebuild the model

Finalize the model

VALIDATION OF MODEL

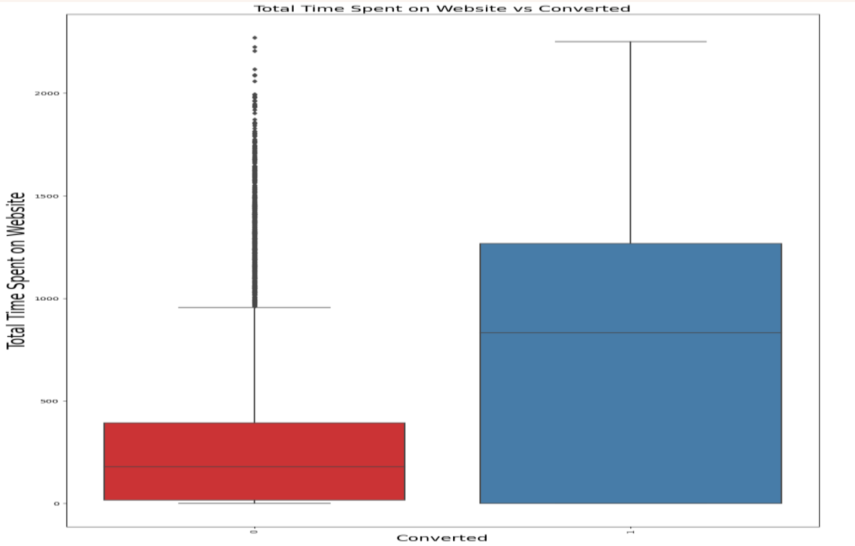
Plotting the ROC Curve

Making predictions on the test set

DATA VISUALIZATION

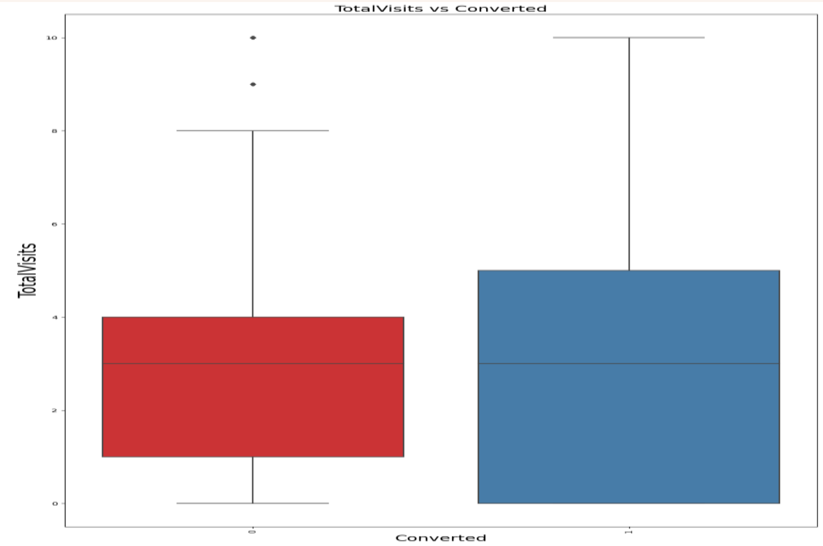
Total time spent on website

Leads spending more time on the Website are more likely to be converted.



Total Visit

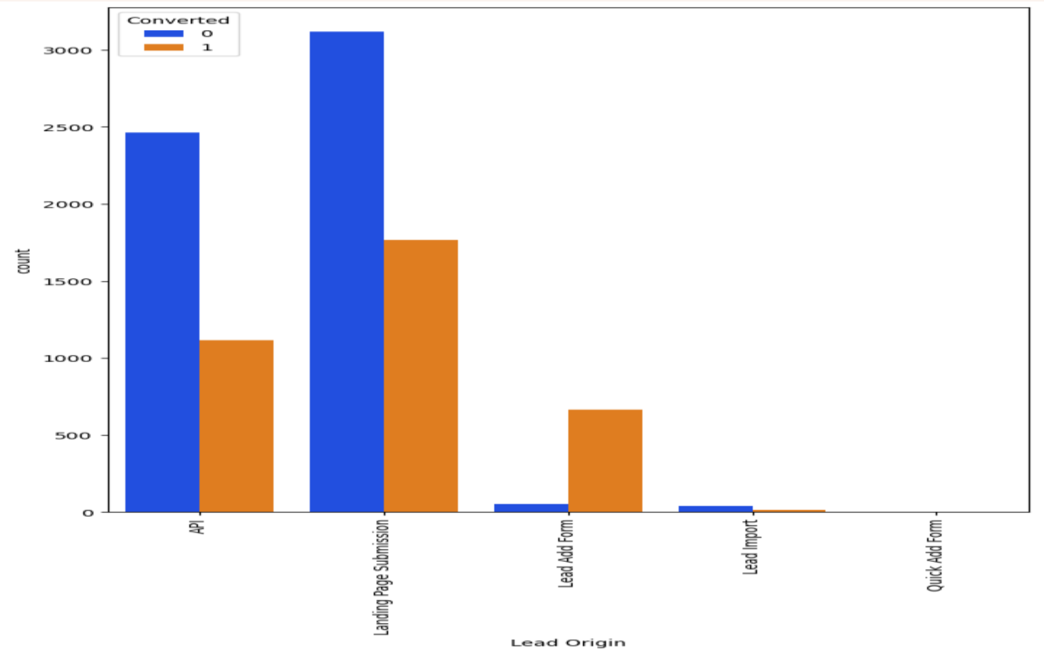
Nothing can be concluded based on Total Visits



EDA

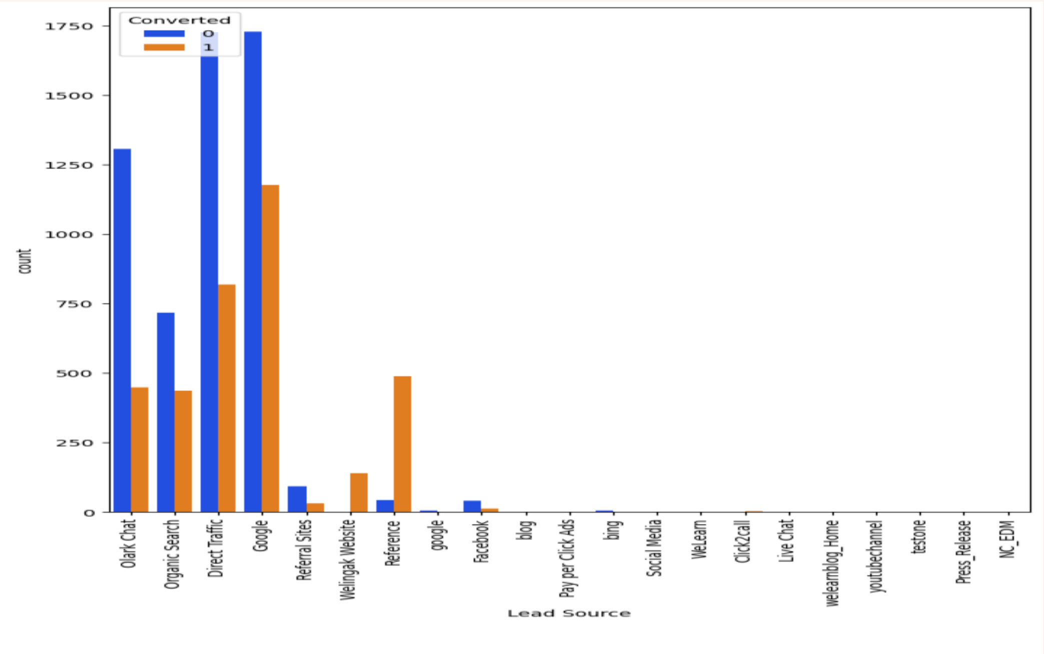
Lead Origin

API and Landing page submission has a smaller number of conversion rate, but they have a greater number of leads



Lead Source

Google and Direct traffic has maximum number of leads.



EDA

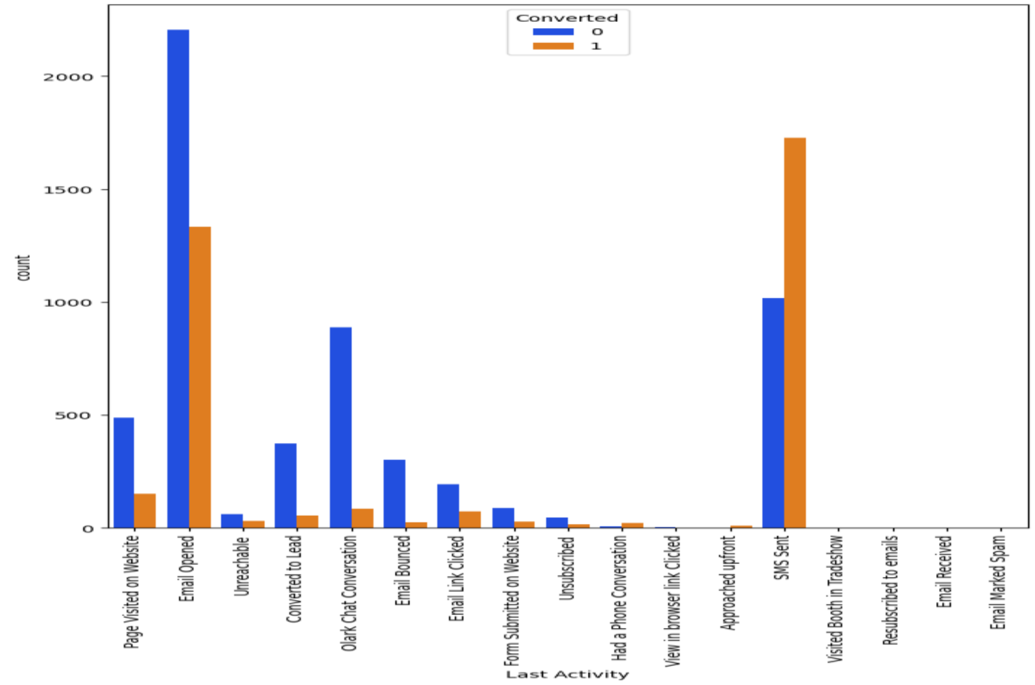
Last Activity

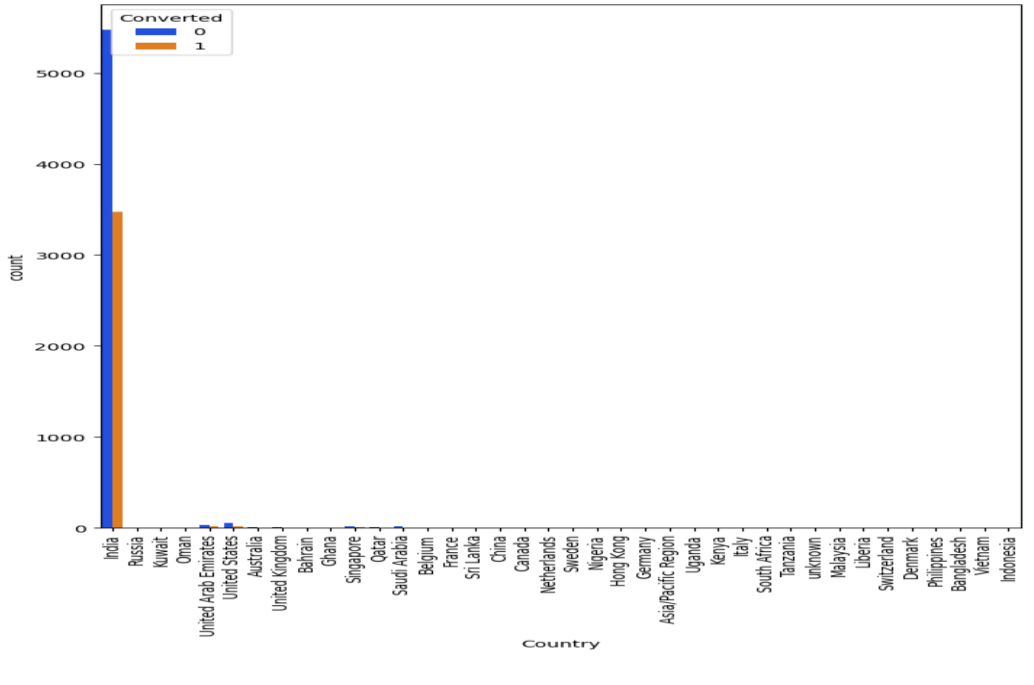
Most of Leads have their last activity Email opened

SMS sent last activity has good conversion rate

Country

We can observe that most of the values belong to India and hence can be dropped





DATA CONVERSION

Numerical Variables Normalized

Dummy Variables created for Object data type

Data Split into Test and Train Set with 70:30 Ratio

Use RFE Feature Scaling

Running RFE Feature Scaling with 15 Variables

Model Building and removing Variable with high p-value and high VIF

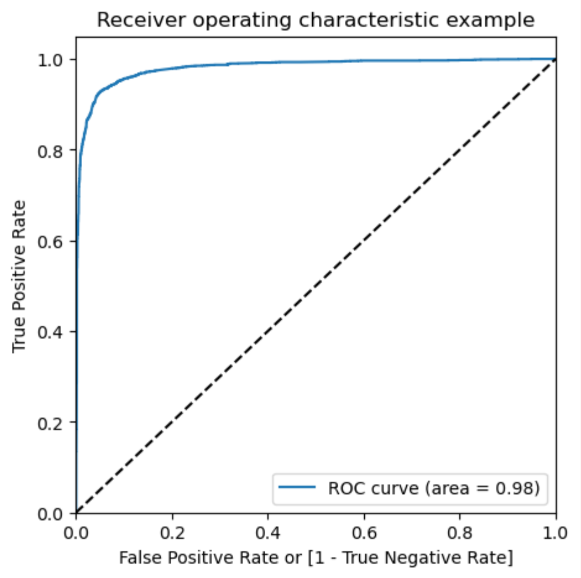
Predictions on test Set

MODEL BUILDING

ROC CURVE

Classifiers that give curves closer to the top-left corner indicate a better performance.

The optimum cut off value in ROC curve is used to find the accuracy, sensitivity and specificity which came to be around >90% each.



CONCLUSION

Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.5 with accuracy, sensitivity of 90% and specificity of 96%

Precision – Recall:

This method was also used to recheck and a cut off 0.5 was found with Precision around 94% and recall around 90% on the test data frame.