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.

- 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%.

- Now, although 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.

Procedure

- Data understanding and exploration

- Data Visualisation

- Outlier Analysis

- Data preparation

- Model building and evaluation

Data understanding and exploration

- Read the csv file

- Analysing columns for null value treatment

- Filtering categorical variables

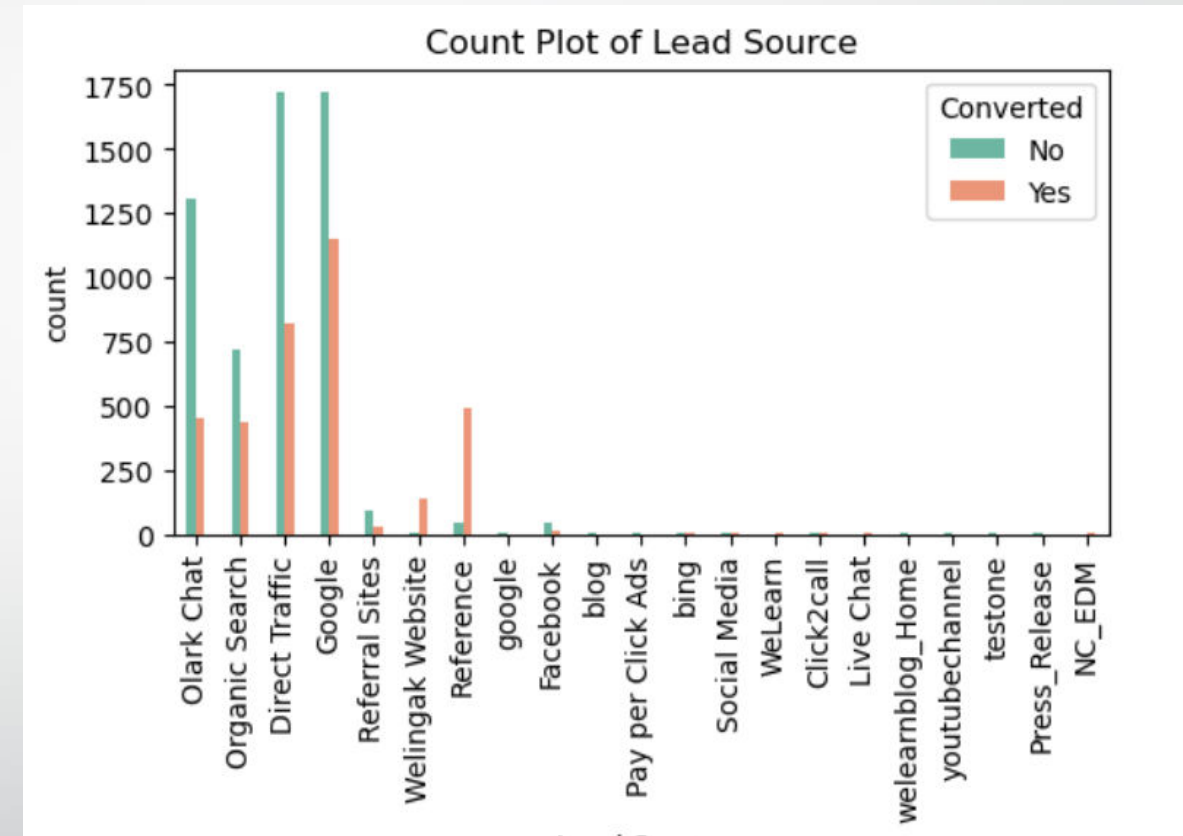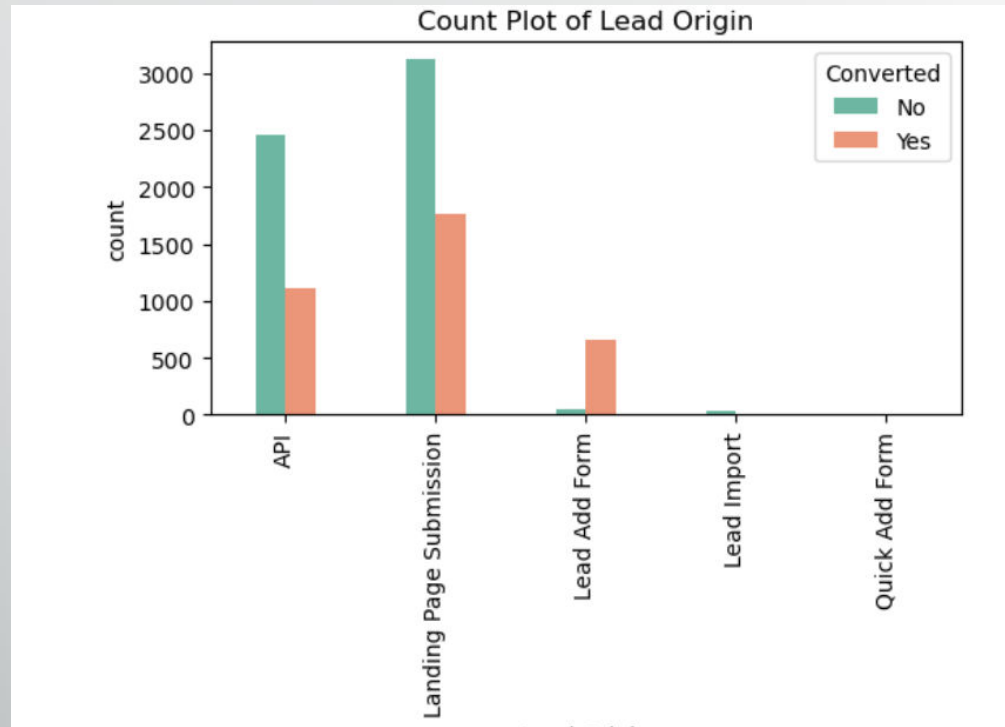- Filtering numerical variables

Outlier Analysis

Histogram of Lead Number

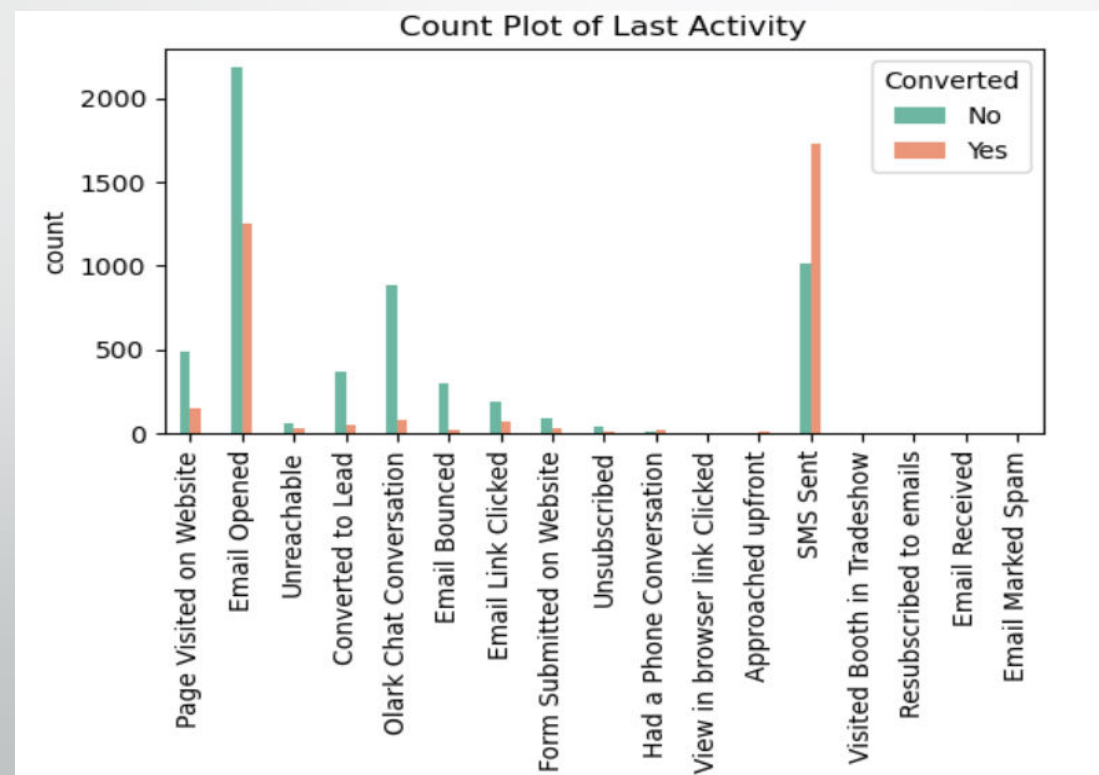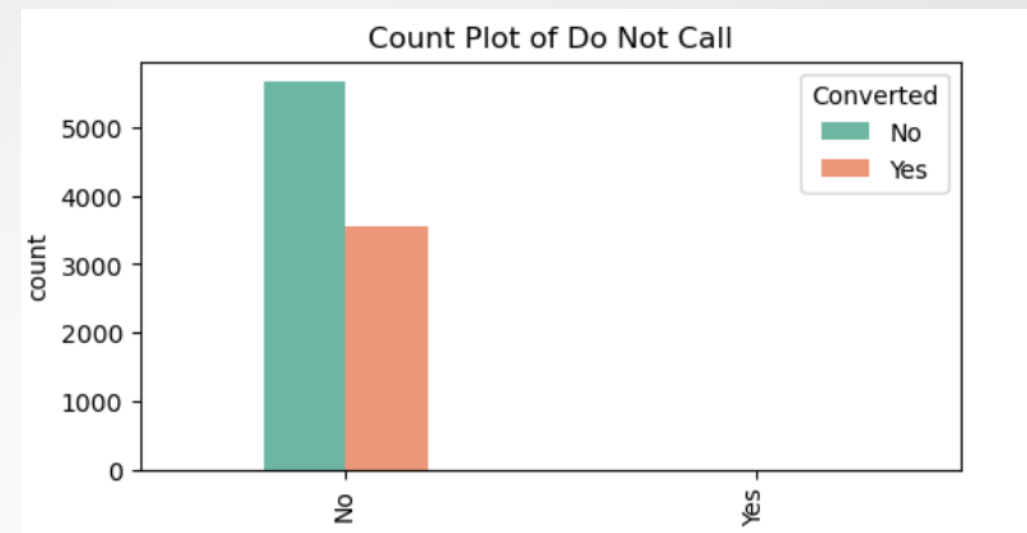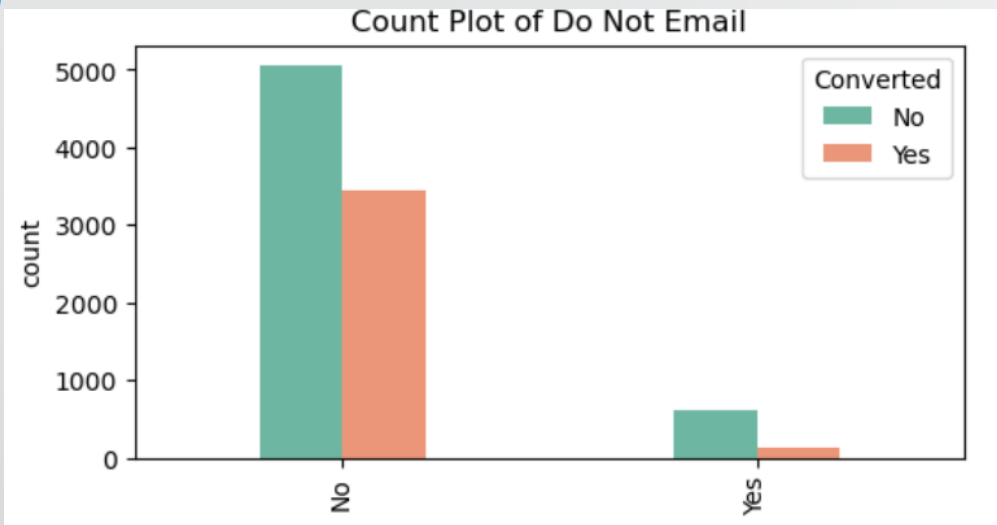Histogram of Total Time Spent on Website
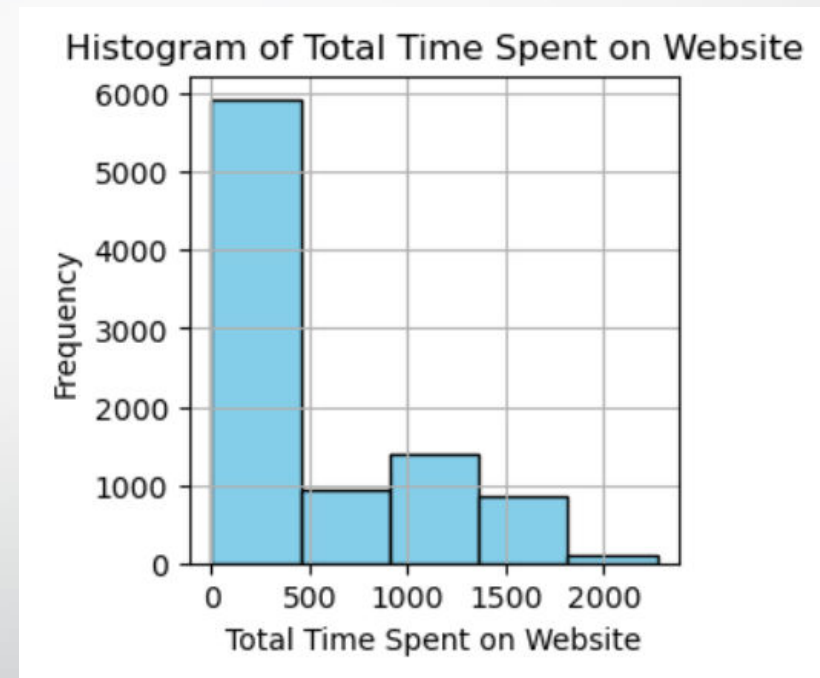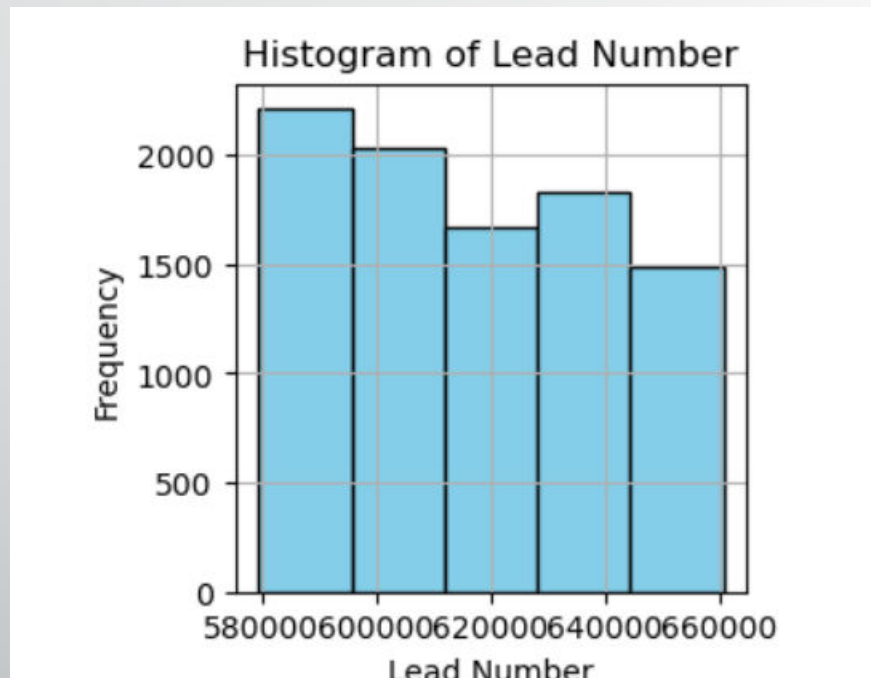
Data preparation

- Columns to be deleted based on the analysis

- Analyzing the final dataframe to be considered for modelling

- Converting binary variables into 0 and 1

- Numerical Variables are Normalised

- Creating dummy variables for categorical values

Categorical Variable Relation

Count Plot of Lead Origin

Count Plot of Lead Source

Count Plot of Do Not Email

Count Plot of Do Not Call


Count Plot of Last Activity

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 .

- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5 .

- Optimal model using Logistic Regression.

- Calculating accuracy, sensitivity, specificity, precision score.

- Checking model ROC Curve

- Checking VIF of the variables

ROC Curve
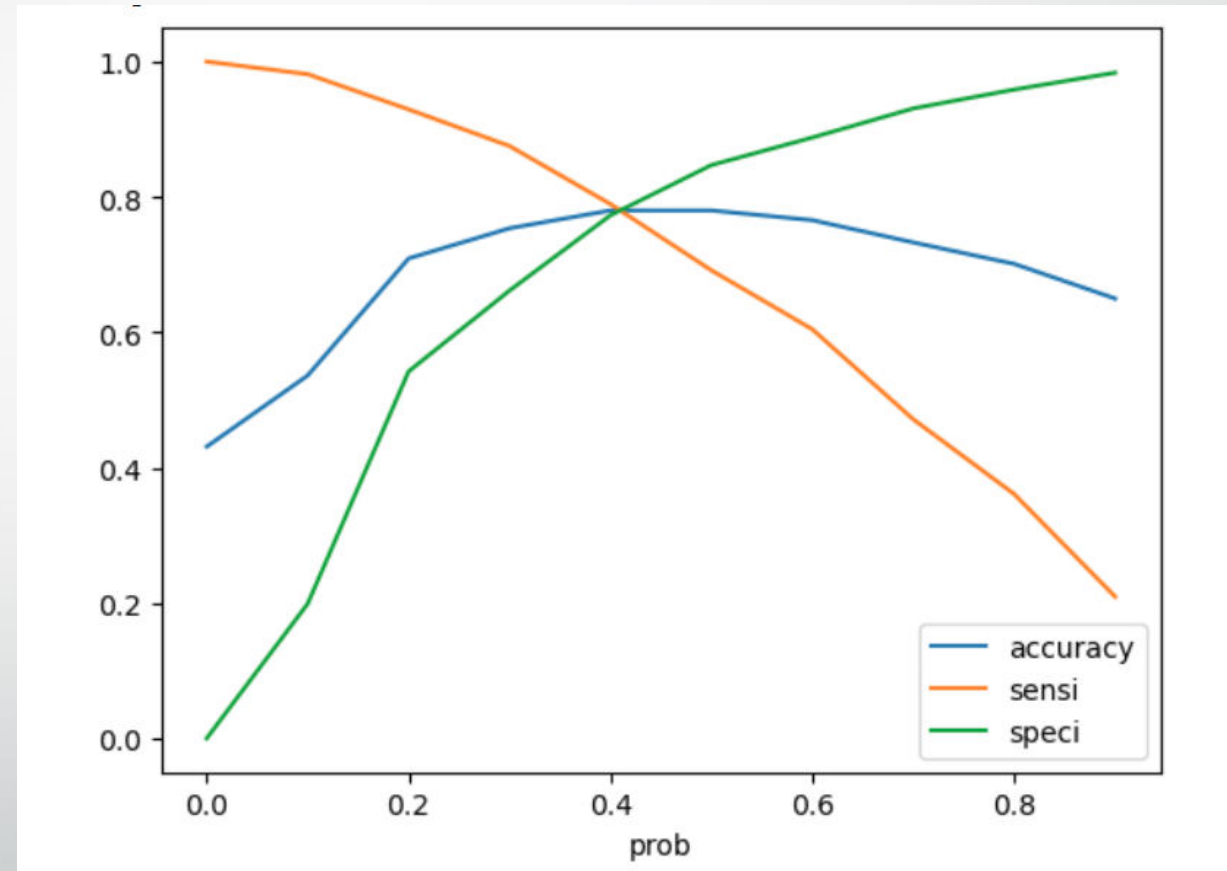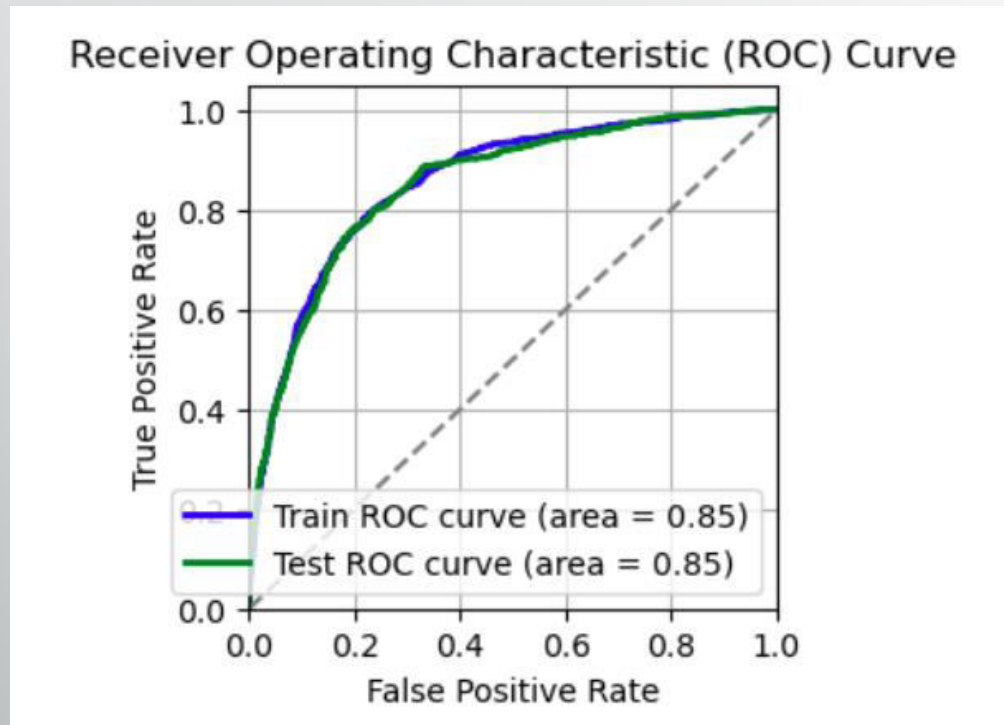
From the second graph it is visible that the optimal cut off is at 0.4

Result

Accuracy Score
Train_Accuracy :- 78.0
Test_Accuracy :- 77.87

Precision Score
Train_precision:- 72.47
Test_precision :- 72.03

Sensitivity
Train_Sensitivity :- 78.94
Test_Sensitivity :- 79.6

Specificity
Train_Specificity :- 77.29
Test_Specificity :- 76.57

Confusion Matrix
[2355   692]
[486   1822]

Test Confusion Matrix
[1000   306]
[202   788]

Conclusion

- ***It was found that the variables that mattered the most are :-***

    1- Total Time Spent on Website

    2- Lead Source_Reference

    3- Last Notable Activity_SMS Sent

    4- Lead Source_Welingak Website

    5- Lead Source_Facebook