Lead Score Case Study

Group Members
1. Shayon Deb
2. Karthikeya M
3. Rahul
4. Preetika Priya

Problem Statement

- X Education sells online courses to industry professionals.

- 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.

Solution Methodology

▶ 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.
3. Multivariate data analysis: Multi-feature impact on the conversion of leads into hot leads

▶ 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.

Data Manipulation

- Total Number of Rows =37, Total Number of Columns =9240.

- Single value features like "Magazine", "Receive More Updates About Our Courses", "Update me on Supply"

- Chain Content", "Get updates on DM Content", "I agree to pay the amount through cheque" etc. have been dropped.

- Removing the "Prospect ID" and "Lead Number" which is not necessary for the analysis.

- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: "Do Not Call", "What matters most to you in choosing course", "Search", "Newspaper Article", "X Education Forums", "Newspaper", "Digital Advertisement" etc.

- Dropping the columns having more than 45% as missing value such as 'How did you hear about X Education' and 'Lead Profile'.
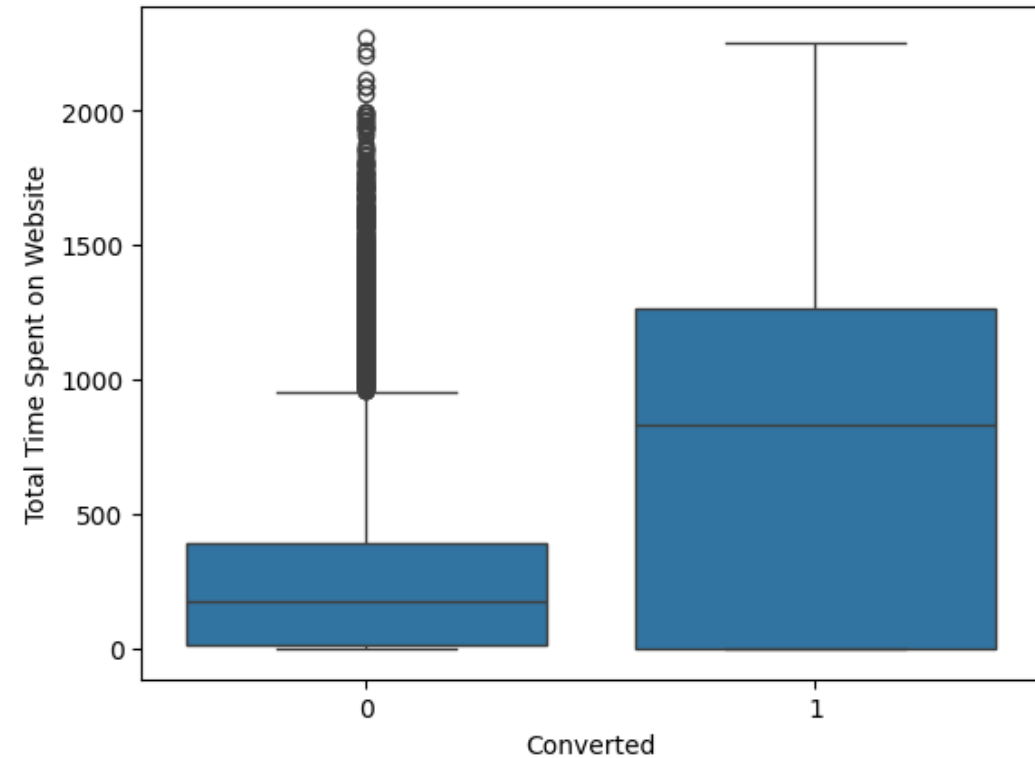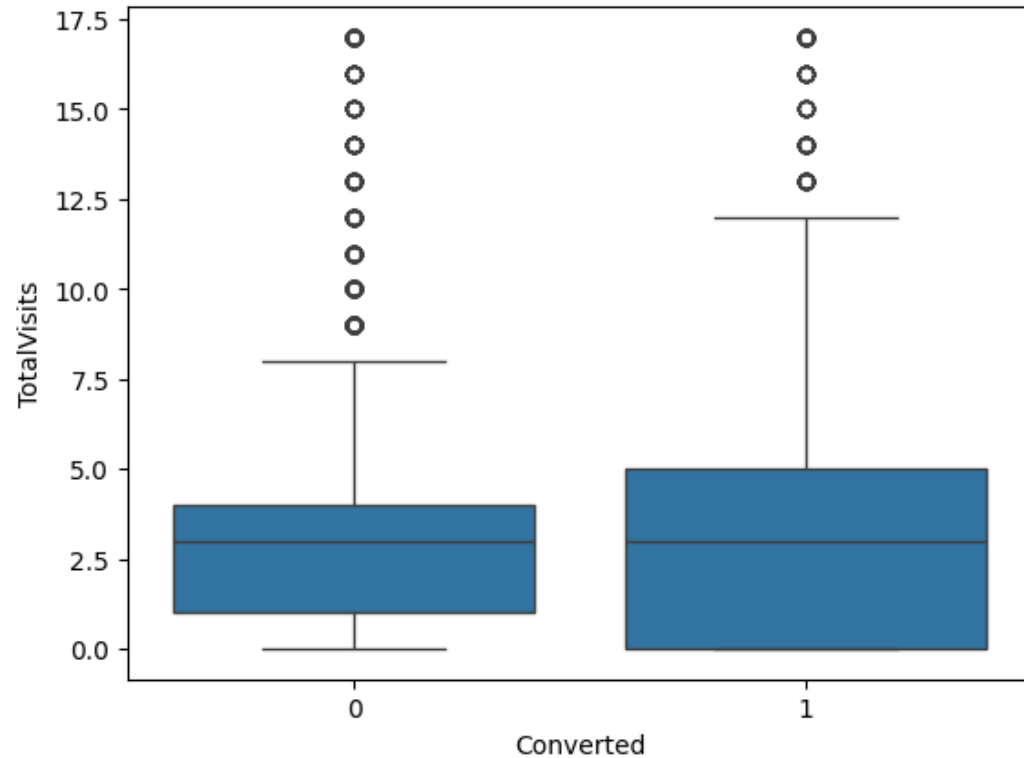
EDA
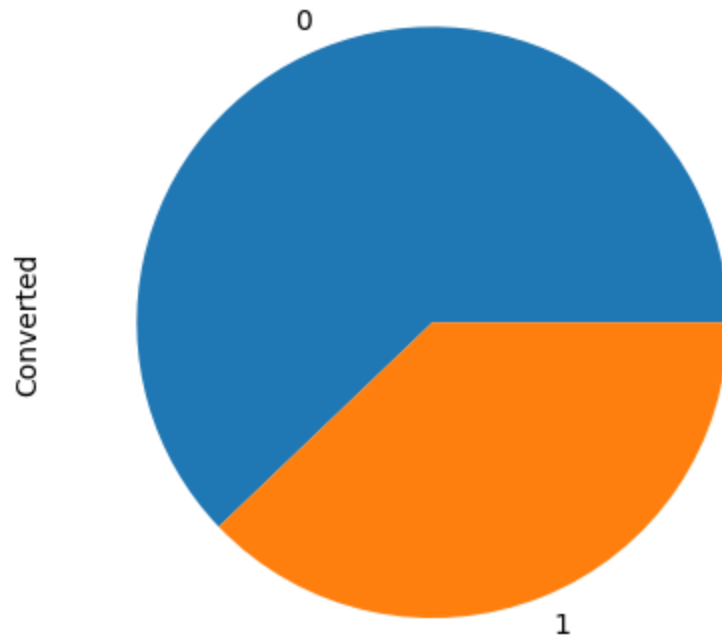
The above pie depicts about 2/3 of the dataset consists of information of cold leads and 1/3 is about the hot leads
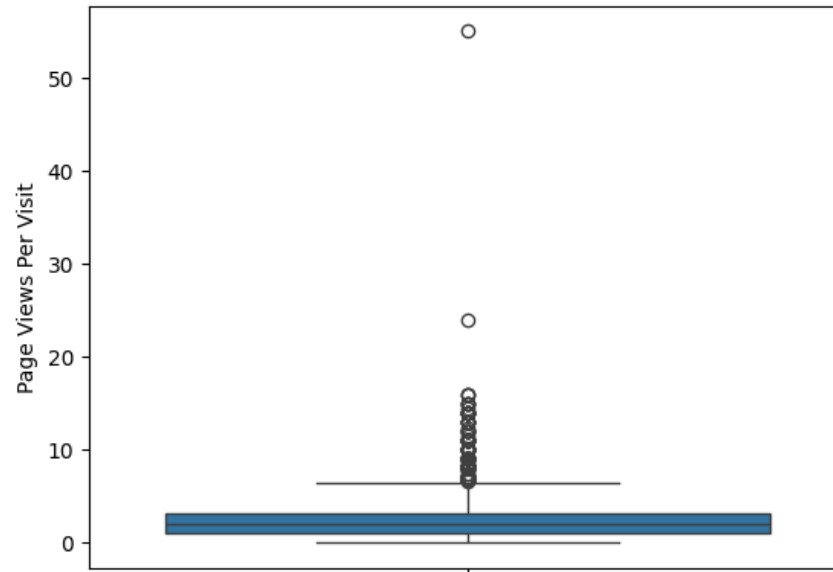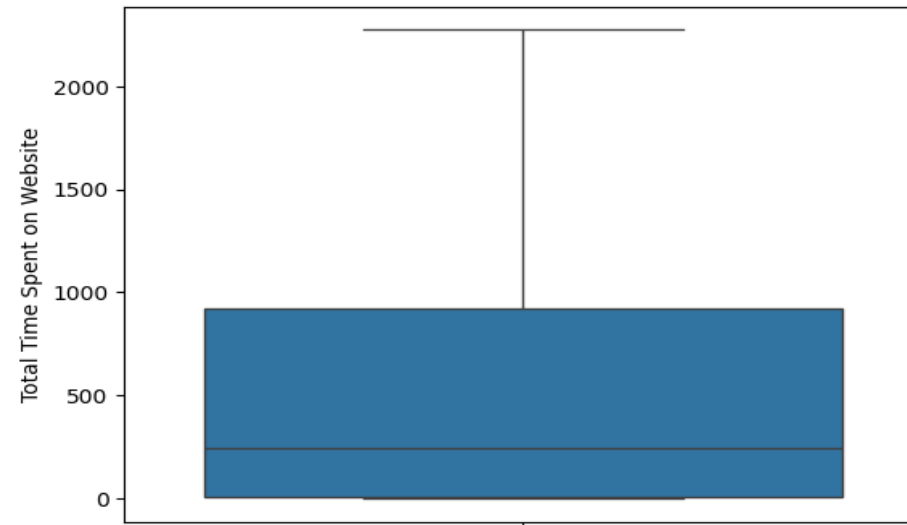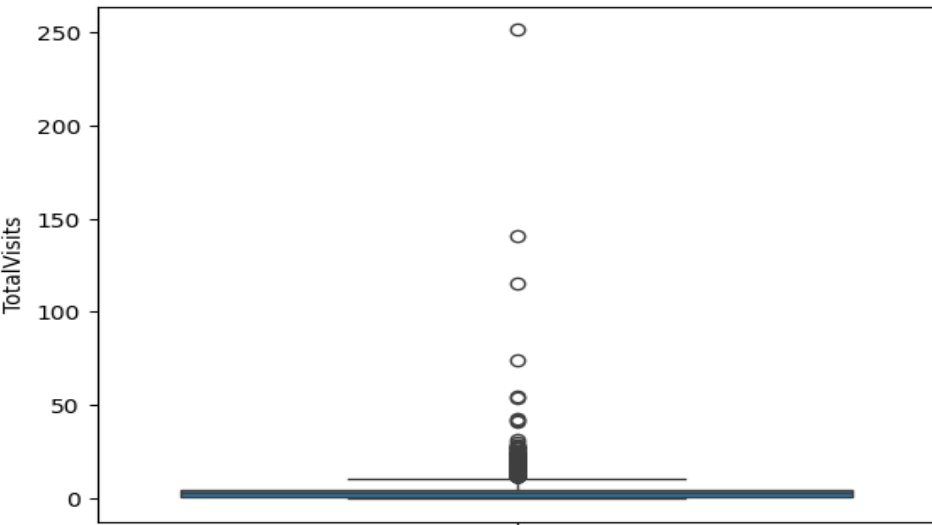
From the above boxplots it is evident that out the all numeric columns 'TotalVists' and 'Page Views Per Visit' column had outliers. These are imputed using the capping method (Used 99th percentile value to impute the data)

Numerical Variable Relation

From the above plots it is evident that the leads who have higher visits and spends more time on the website are likely to be the hot leads
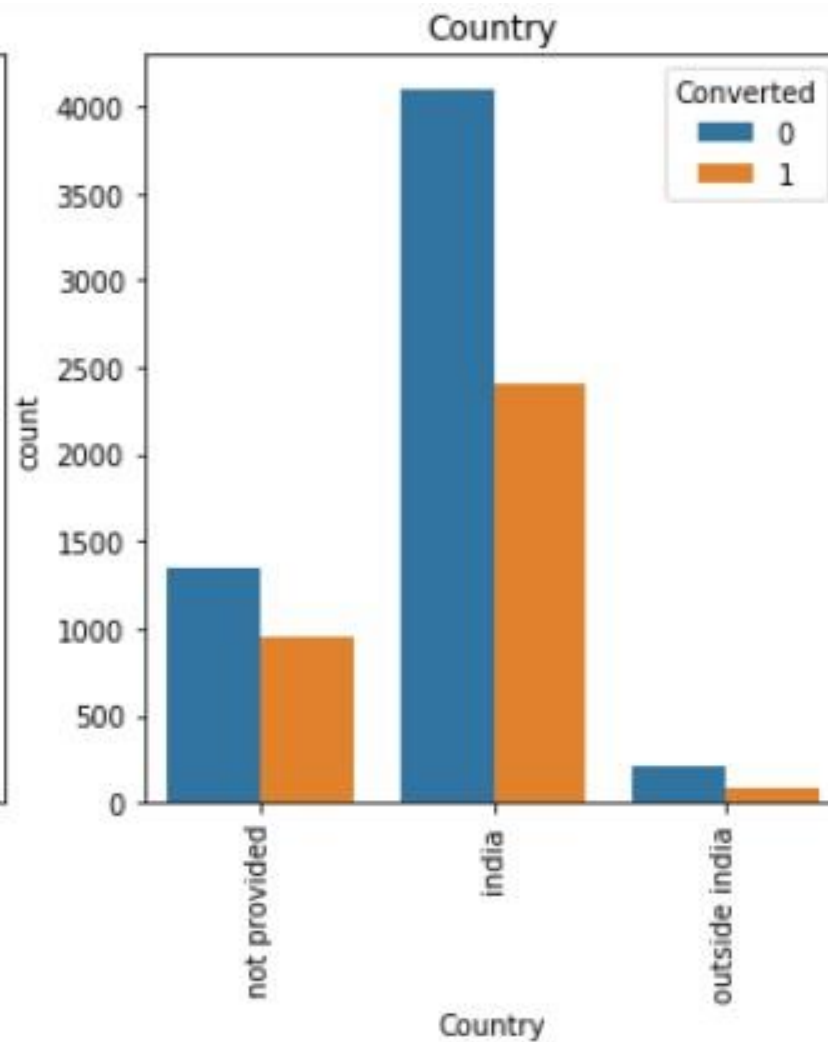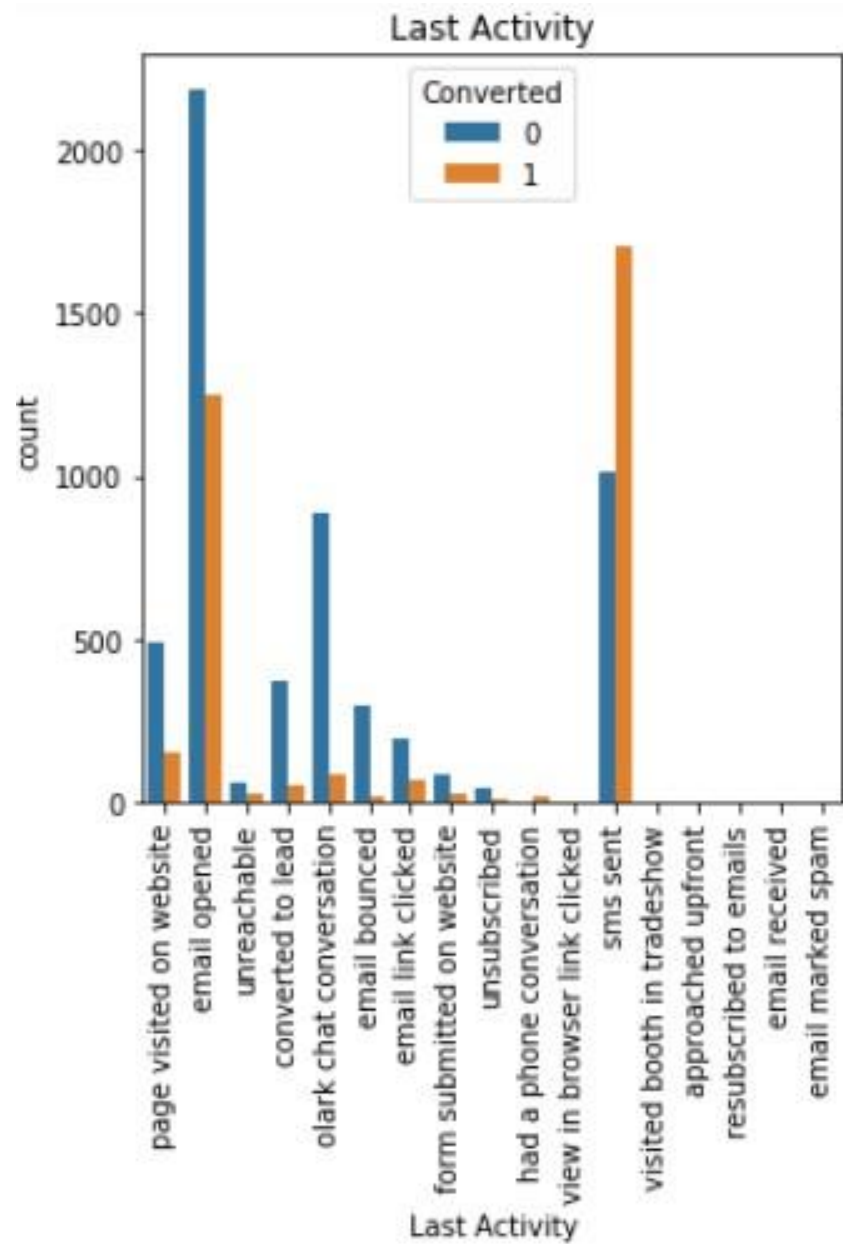
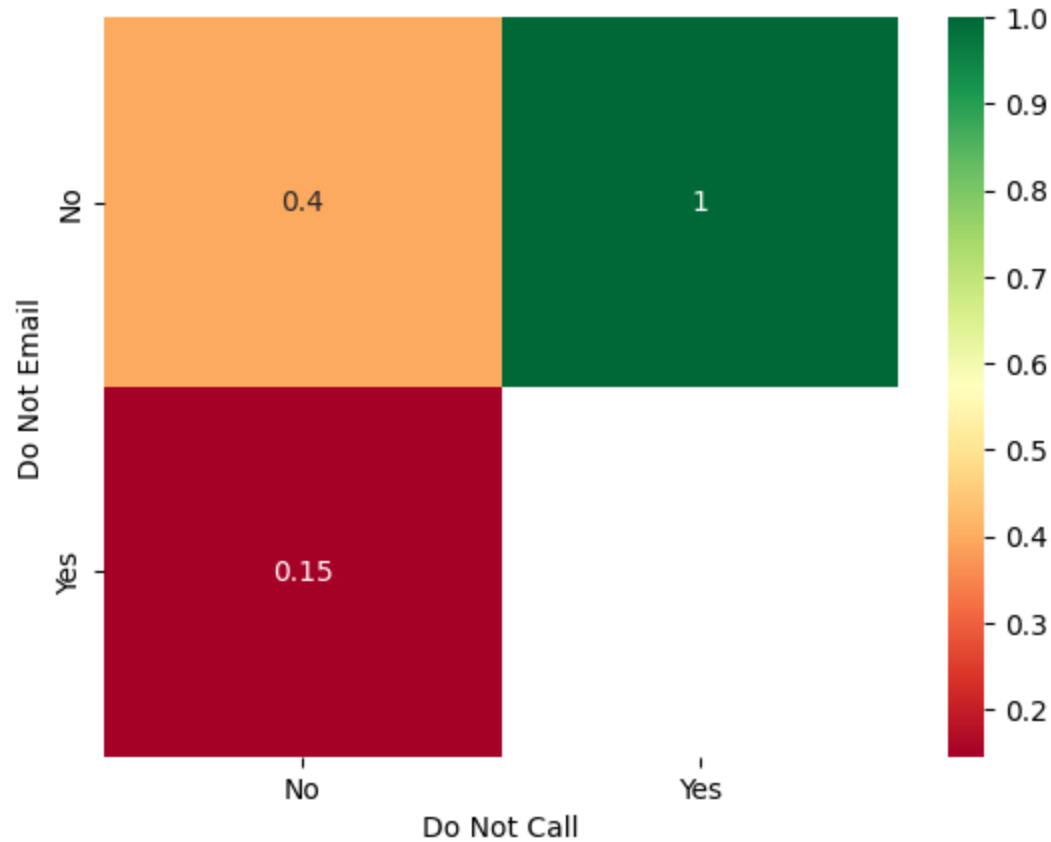# Categorical Variable Relation
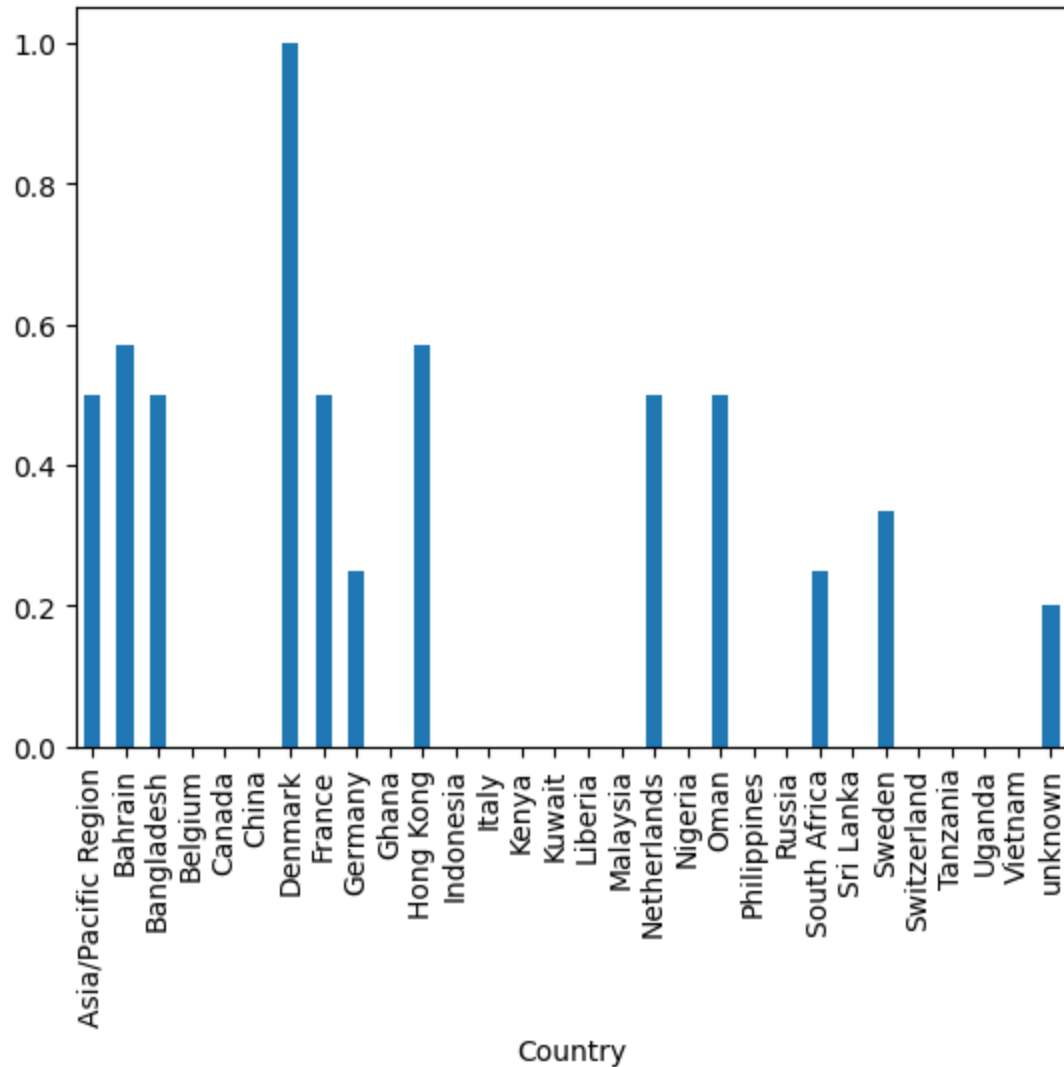
So, From the graph it is very clear that , there are very Few leads(<10 leads) from these and most of these countries even have conversion ratio as zero. So, it is better not to spend much efforts on these countries
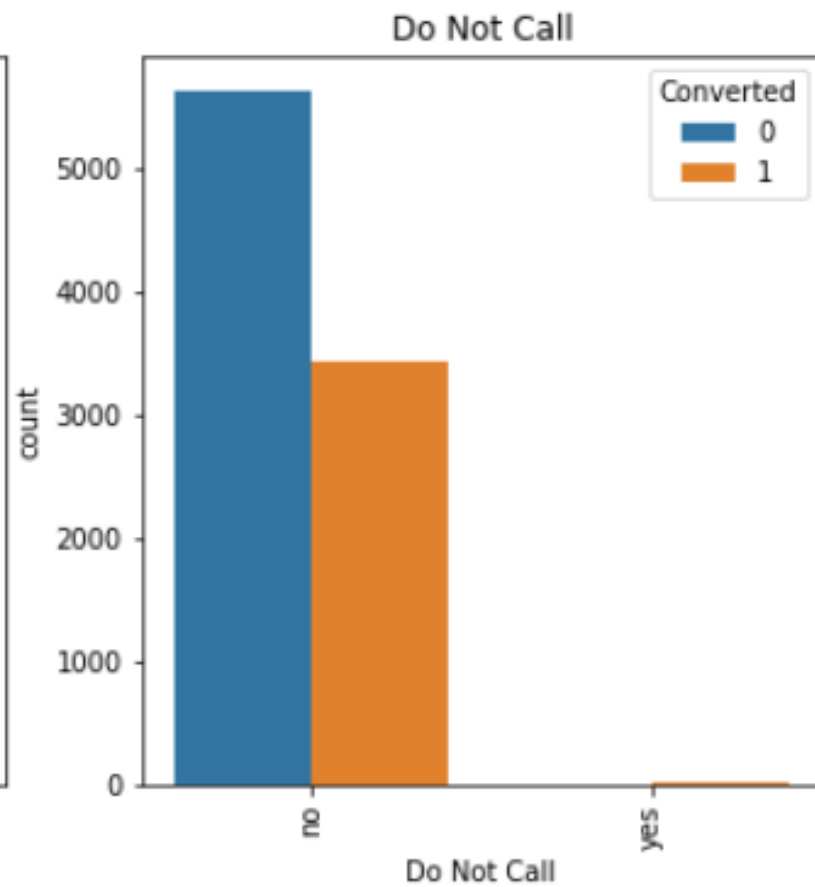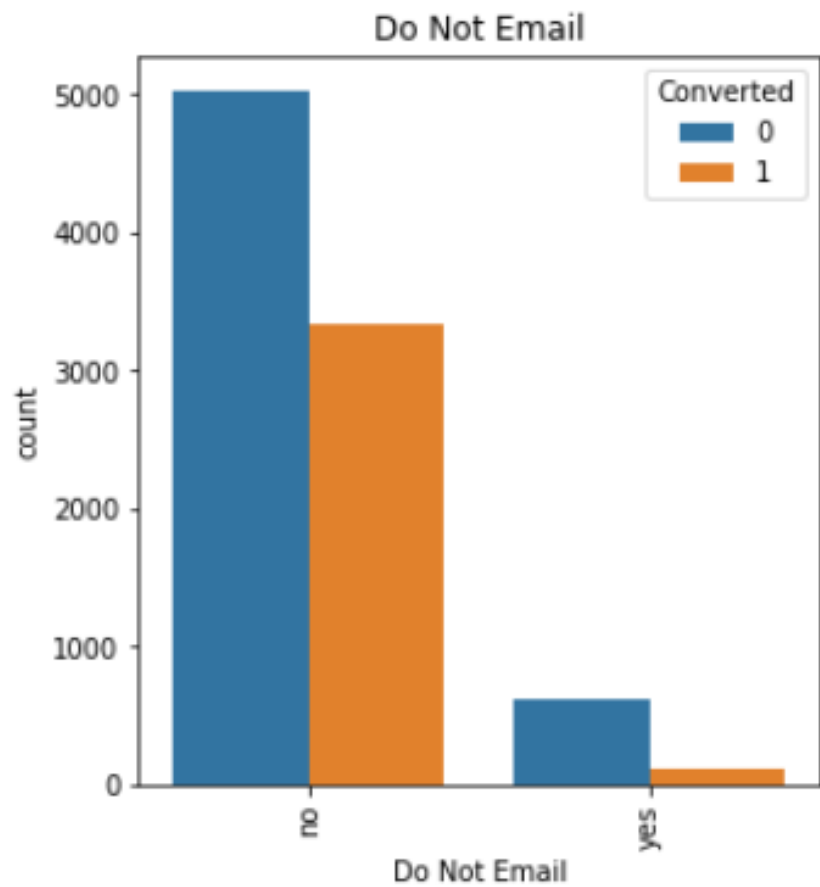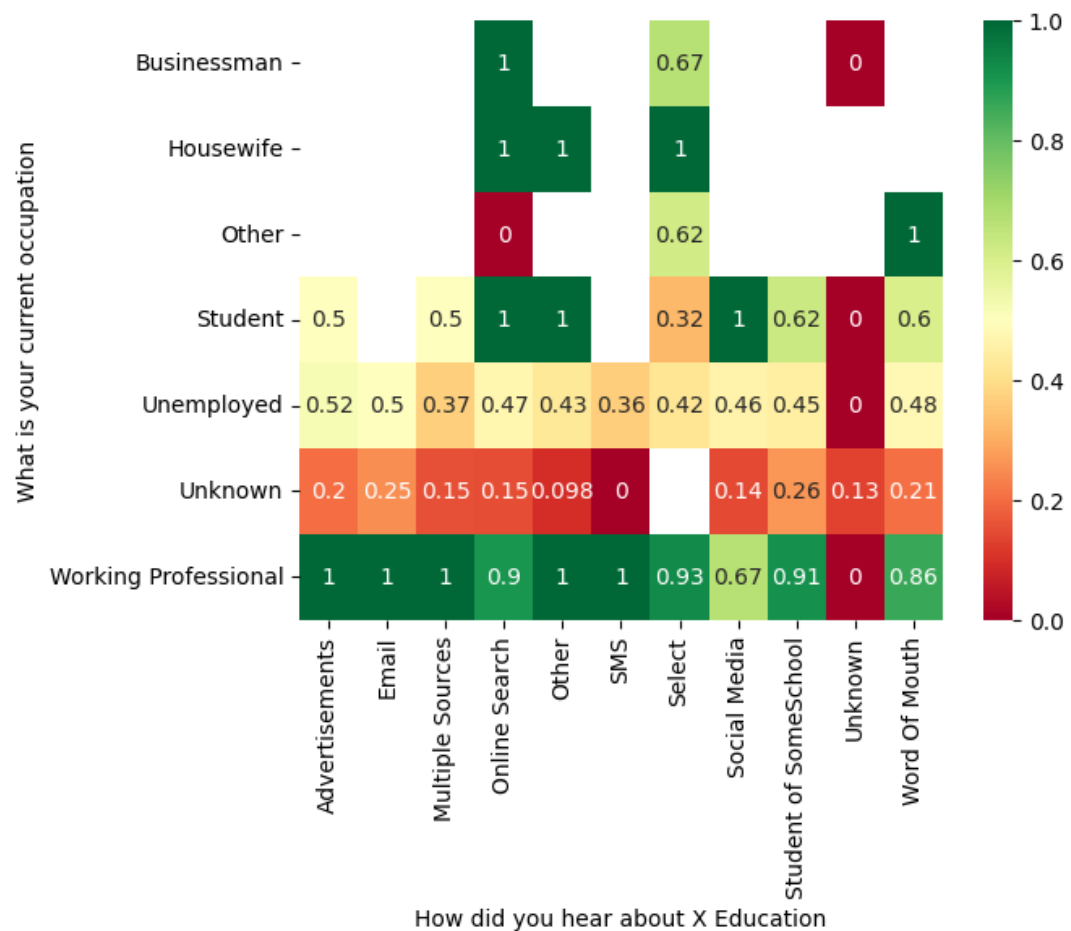
Mutli-variate analysis

The graph is created using both the categorical columns 'Do Not Email' and 'Do Not Call' with average of 'Converted values

The graph depicts that leads who do not want to receive calls but want receive mails have the highest percentage of the conversion into hot leads

The graph is created using both the categorical columns 'What is your current occupation' and 'How did you hear about X Education' with average of 'Converted values

From the above it is clear that the working professionals are the ones who are more likely to get converted out of which the conversion rate is very low from the social media promotions to the working professionals
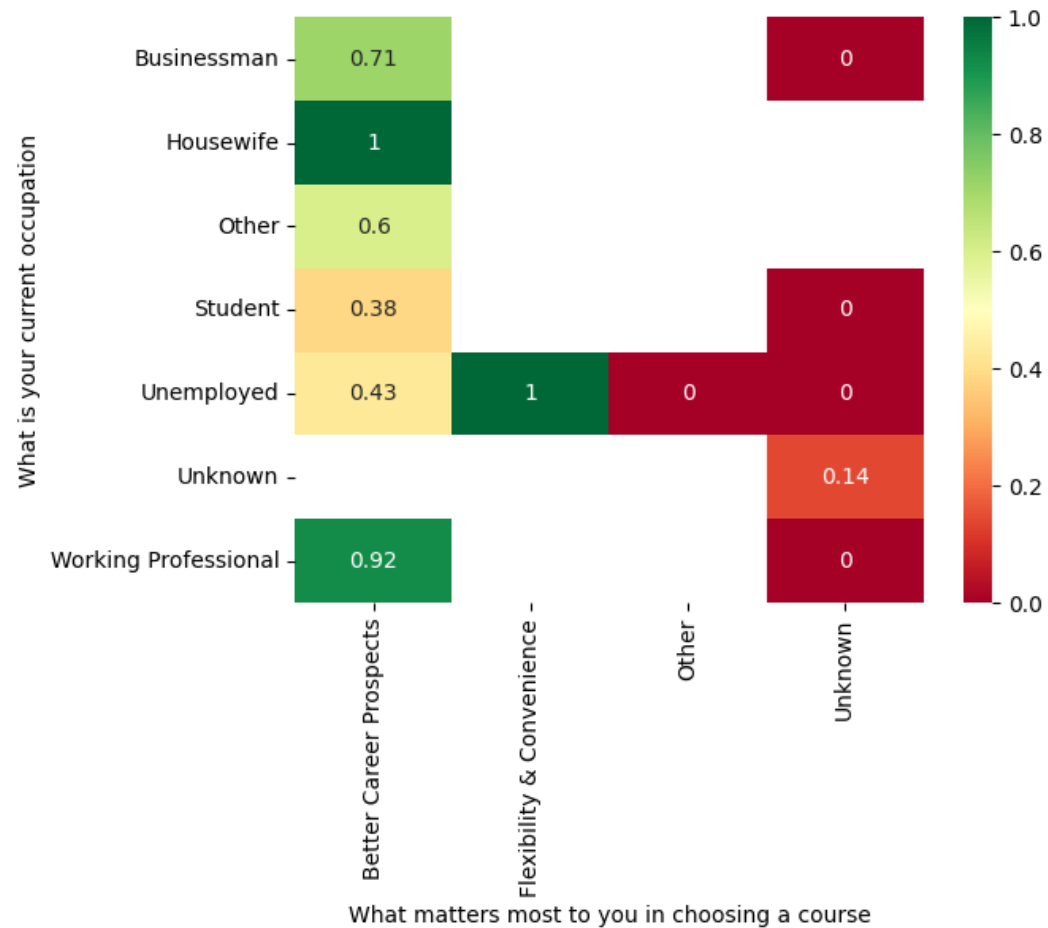
The graph is created using both the categorical columns 'What is your current occupation' and 'What matters most to you in Choosing a Course' with average of 'Converted values

From the plot, it is clear that both working professionals and housewives who are choosing course for better career prospects ,the unemployed choosing Flexibility and convenience as most mattered aspects are likely to be the hot leads

Data Conversion

- ► Numerical Variables are Normalised

- ► Label encoder used for object type variables

- ► Total Rows for Analysis: 9074

- ► Total Columns for Analysis: 27

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 15 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 81%

ROC Curve
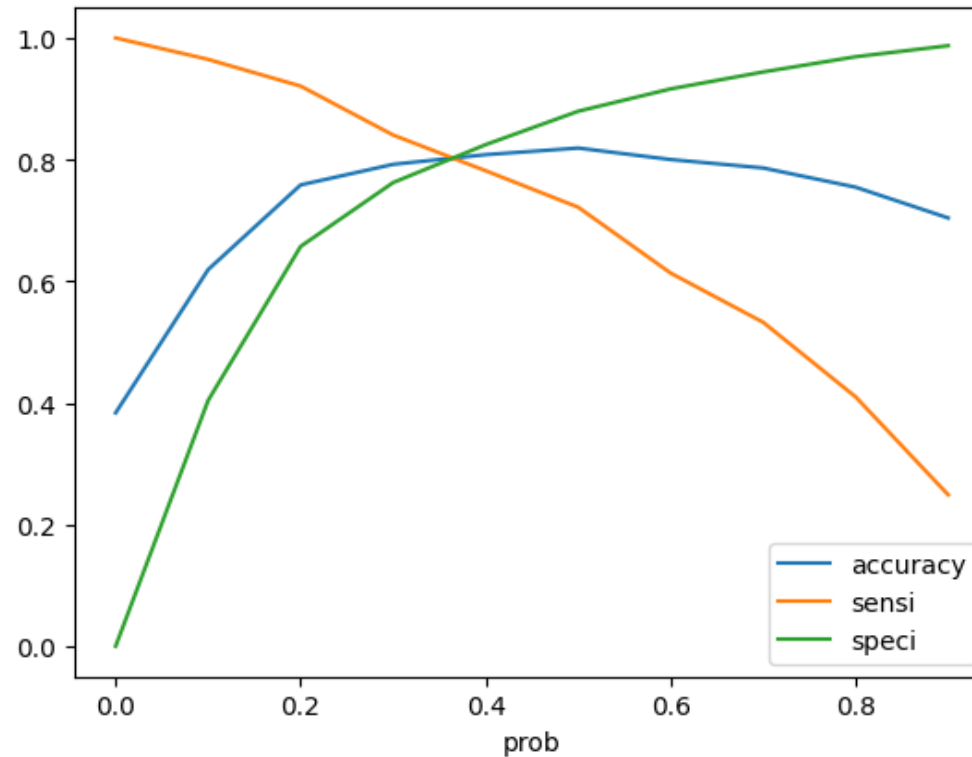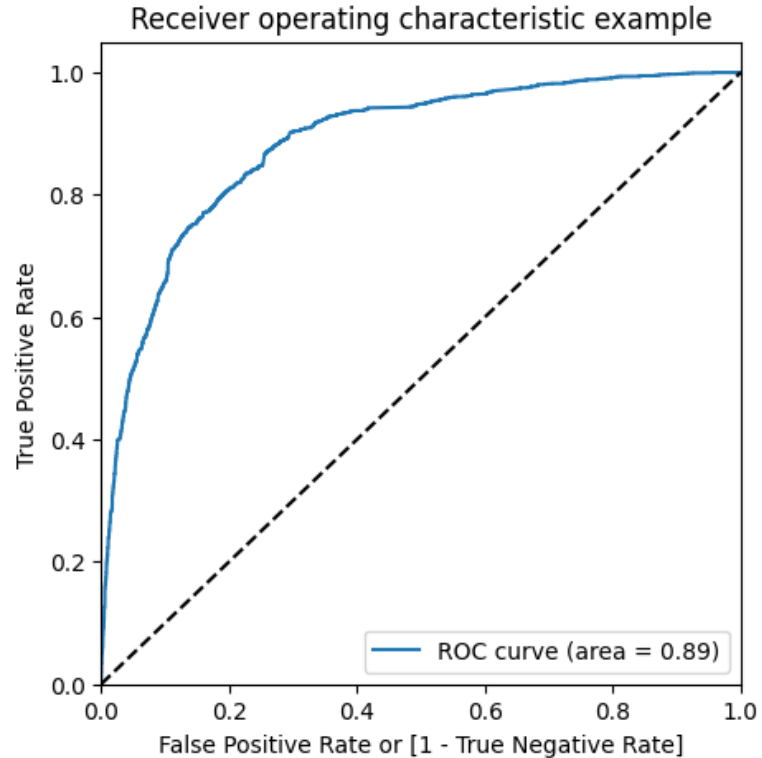
- **Finding Optimal Cut off Point**

- Optimal cut off probability is that

- probability where we get balanced sensitivity and specificity.

- From the second graph it is visible that the optimal cut off at 0.425.

Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

► The total time spend on the Website.

► Total number of visits.

► When the lead source was:
a. Google
b. Direct traffic
c. Organic search
d. Welingak website

► When the last activity was:
a. SMS
b. Olark chat conversation

► When the lead origin is Lead add format.

► When their current occupation is as a working professional.
Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.