LEAD SCORING CASE STUDY

Batch: DS C63 Batch
Submitted By:
Subhashree Chiranjita Baral
Caroline Abraham
Surashree Chakrabarty

Problem Statement:

X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once people land on the website, they might browse the courses, fill out a form for a course, or watch some videos. When they provide their email address or phone number through a form, they are classified as leads. Additionally, the company acquires leads through past referrals. Once leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some leads get converted, while most do not. The typical lead conversion rate at X Education is around 30%.

Business Objectives:

X Education needs assistance in selecting the most promising leads – those most likely to convert into paying customers. The company requires a model that assigns a lead score to each lead, indicating a higher likelihood of conversion for leads with higher scores (hot leads) and a lower likelihood for those with lower scores (cold leads). The CEO has specified a target lead conversion rate of approximately 80%

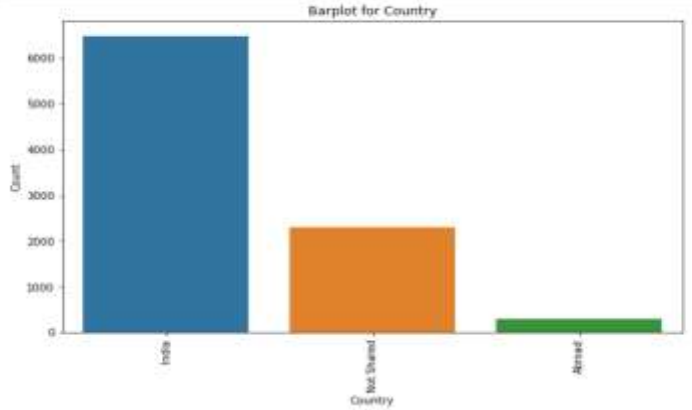
Steps Taken:

1. Data Sourcing
2. Data loading and Data Cleaning :
   - Imputing and Removing Missing columns
   - Handling Outliers
3. Exploratory Data Analysis
   - Univariate Analysis
   - Bivariate Analysis
   - Multivariate Analysis
4. Data Preparation:
   - Convert binary variables have been into 0's and 1's
   - Get dummies
5. Train Test Split:
6. Feature Scaling
7. Building Model:
   - Building a model using RFE approach
   - Building the model using features selected by RFE
   - Calculating VIFs
   - Plotting Heatmap to represent the relationship between selected columns
   - Repeating steps 2-4 of building the model until all VIFs are below 5 and all p-values are below 0.05Model Evaluation Sensitivity and Specificity on Train Data Set
8. Model Evaluation - Precision and Recall on Train Data Set
9. Model Evaluation - Precision and Recall on Test Data Set

Univariate Analysis:

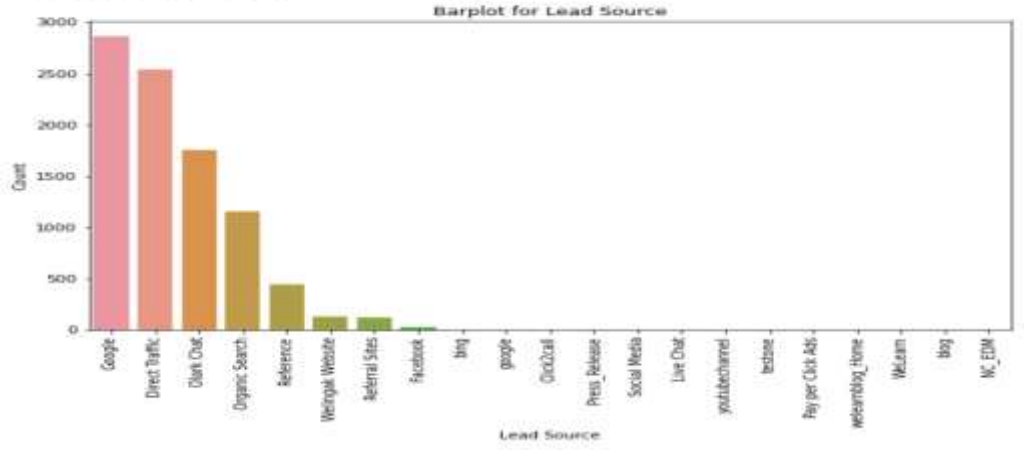
Univariate analysis is a statistical method used to analyze and describe the distribution of a single variable. The goal is to summarize and describe the central tendency, dispersion, and shape of the distribution of the variable.

```
In [496]: 1 graph_cat_cols(df['Country'])
```



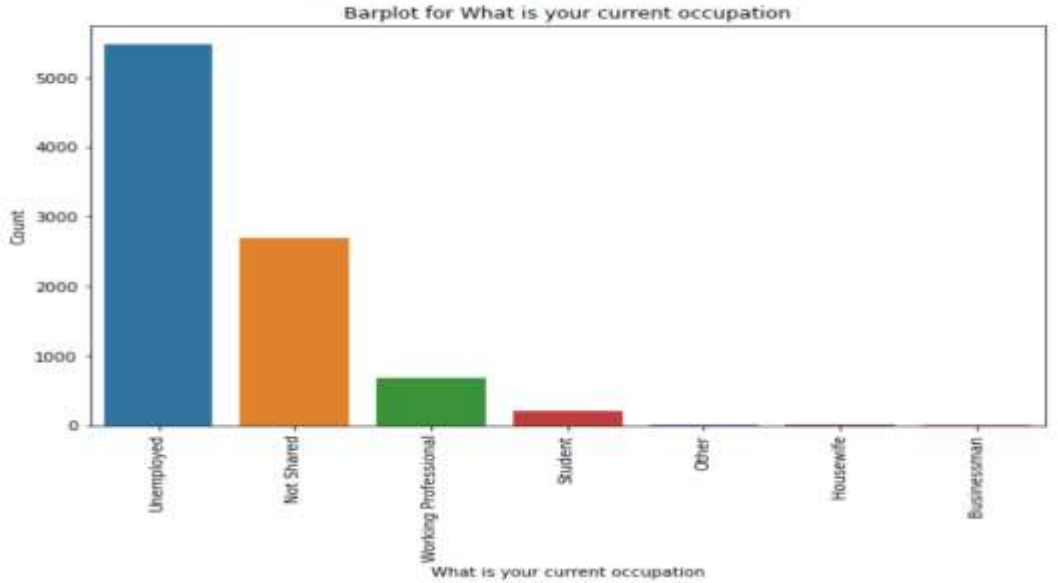
Insights/Assumptions:
The analysis reveals that most of the leads are from India

```
1 # Plotting graph for 'Lead Source'
2 graph_cat_cols(df['Lead Source'])
```



Insights/Assumptions:
We can see that most of the lead's lead source is Google followed by Direct Traffic

```
# Plotting graph for 'What is your current occupation
graph_cat_cols(df['What is your current occupation'])
```



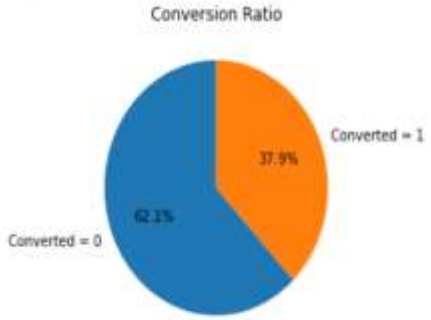
Insights/Assumptions:

Here, we can see that most of the leads are unemployed

```
target_counts = df['Converted'].value_counts()

# Create a pie chart

plt.pie(target_counts, labels=["Converted = 0","Converted = 1"], autopct='%1.1f%%', startangle=90)
plt.title('Conversion Ratio')
plt.show()
```



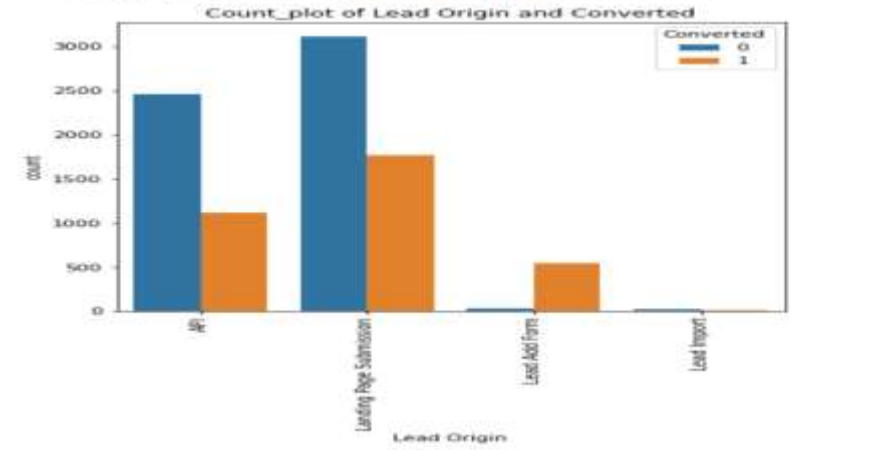
Insights/Assumptions:

We can see that the conversion rate is 37.9%.

Bivariate Analysis:

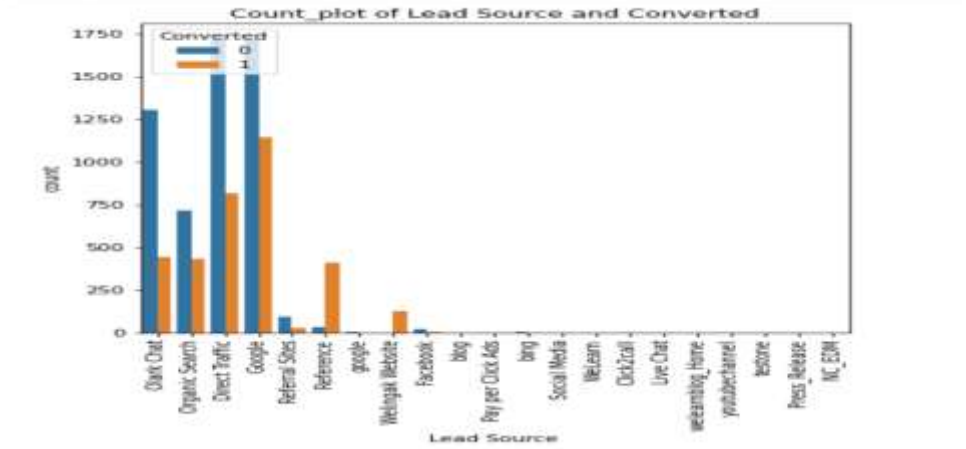
Bivariate analysis is a statistical method used to examine the relationship between two variables. It involves analysing the joint distribution of two variables to understand how they are related to each other. The primary goal of bivariate analysis is to determine whether there is a statistically significant association or correlation between the two variables.

```
bi_graphs_num_cat('Lead Origin')
```



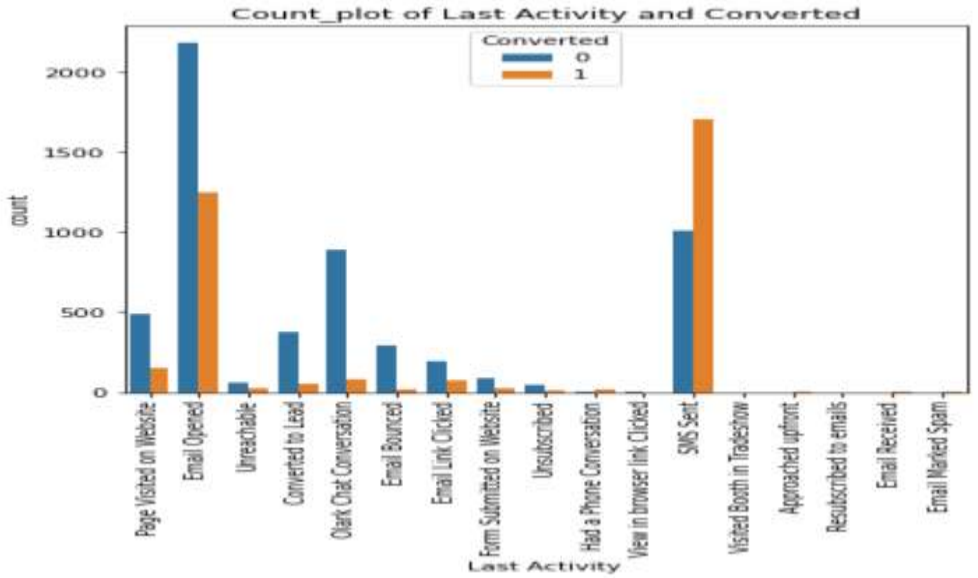
Insights/Assumption:
Leads having 'Lead Origin as 'Lead Add form' are more converted leads

```
bi_graphs_num_cat('Lead Source')
```



Insights/Assumptions:
We can see that Google and Direct Traffic highest number of leads but for most of them the converted factor is 0 which means they are not converted. But actually 'Reference' and 'Welingak Website' have more converted leads

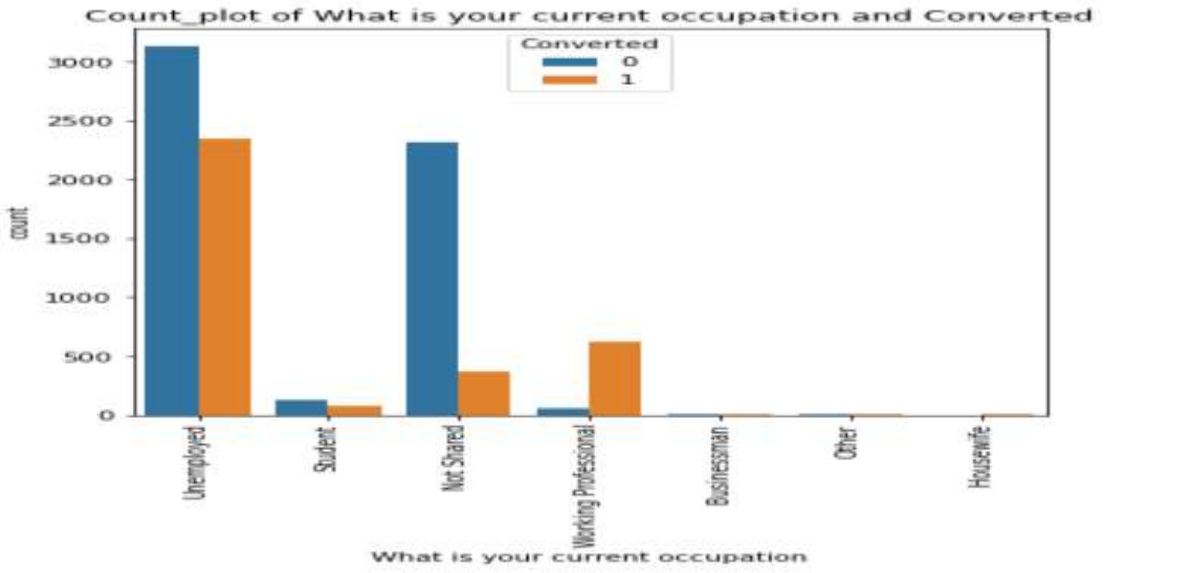
```
1 bi_graphs_num_cat('Last Activity')
```



Insights/Assumptions:

Leads with Last Activity as SMS Sent has more converted leads

```
1 bi_graphs_num_cat('What is your current occupation')
```



Insights/Assumptions:

Working Professional Conversion rate is positive as most of them do convert to leads

Multivariate Analysis:

Multivariate analysis involves the simultaneous analysis of three or more variables to understand complex relationships and patterns in data.

```
#Setting the figure size
plt.figure(figsize=[10,10])
#Plotting the heatmap
sns.heatmap(df[num_cols].corr(),annot=True)
plt.show()
```


```
#Printing the corelation
df[num_cols].corr()
```

	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Converted	1.000000	0.041091	0.359261	0.001435
TotalVisits	0.041091	1.000000	0.296085	0.656896
Total Time Spent on Website	0.359261	0.296085	1.000000	0.330139
Page Views Per Visit	0.001435	0.656896	0.330139	1.000000

Insights/Assumptions:

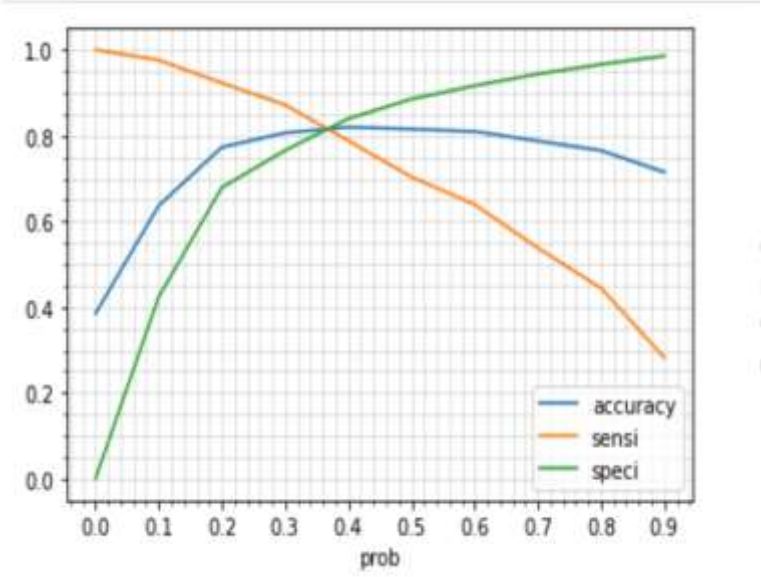
We can see that the target variable is 'Total Time Spent on Website' and 'TotalVisits'

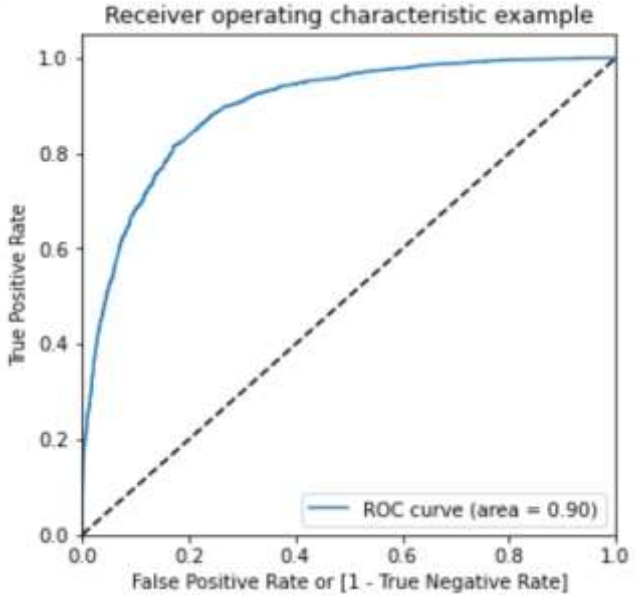
Variables Impacting the Conversion Rate

- Lead Origin-Lead Add Form
- Lead Source-Welingak Website
- Last Activity-Email Bounced
- Last Activity-Had a Phone Conversation
- Last Activity-Olark Chat Conversation
- Last Activity-SMS Sent
- What is your current occupation-Working Professional
- Last Notable Activity-Email Link Clicked
- Last Notable Activity-Email Opened
- Last Notable Activity-Modified
- Last Notable Activity-Olark Chat Conversation
- Last Notable Activity-Page Visited on Website
- Last Notable Activity-Unreachable
- Do Not Email
- TotalVisits
- Total Time Spent on Website
- Page Views Per Visit

Model Evaluation - Sensitivity and Specificity on Train Data Set

- The graph depicts an optimal cut-off of 0.36 based on Accuracy, Sensitivity, and Specificity
- The Area under ROC curve is 0.90 which is a good ROC
- Overall accuracy: 81%
- Sensitivity is 82%
- Specificity 81%
- False Positive Rate 18%
- Positive predictive value 73%
- Negative predictive value 88%



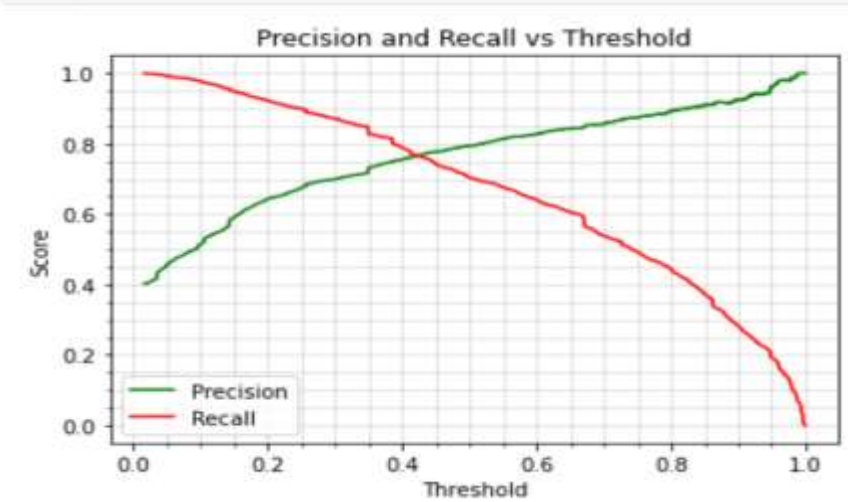


Confusion Matrix

3182	723
429	2017

Model Evaluation - Precision and Recall on Train Data Set

- The graph depicts an optimal cut-off of 0.41 based on Precision and Recall
- Overall accuracy: 82%
- Precision is 76%
- Recall 78%

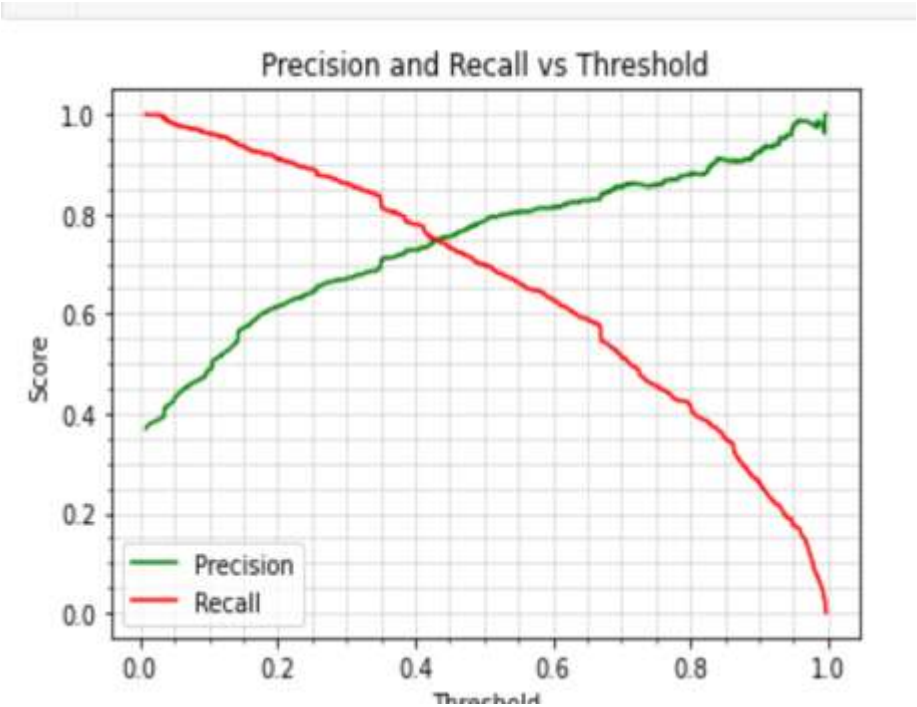


Confusion Matrix

3303	602
537	1909

Model Evaluation – Precision and Recall on Test Dataset

- Overall accuracy: 81%
- Precision is 73%
- Recall 77%
- Sensitivity is 77%
- Specificity 83%



Confusion Matrix

1453	281
222	767

Key Conclusions:

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut-off based on Precision and Recall for calculating the final prediction.
- Lead Origin: Lead Add Form
- 'What is your current occupation: Working Professional
- When the 'lead source' was: the Welingak website
- When the 'last activity' was:
- a. SMS
- b. Olark chat conversation
- c. Had a Phone Conversation
- 'TotalVisits'
- 'Total Time Spent on Website
- 'Page Views Per Visit'
- The precision-recall on the train set is 76% and 78% which are close to the precision and recall of the test set
- The Sensitivity and Specificity of the train set are 82% and 81%