Lead Score Study Presentation

AUGUST 2024

agenda

- Executive summary
- Steps involved
- Understanding the data
- Imputation
- Univariant, bivariant and multivariant analysis
- Model building
- Conclusion/insights
- Recommendations

Executive Summary

X Education company sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. When these people fill up a form providing their email address or phone number, they are classified to be a lead.

X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires to build a model wherein lead score is assigned to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

In this task, we understand how different features influence the conversion rate using Logistic regression.

Steps Involved

1. Understand the problem statement
2. Understanding the data.
    A. Get insight of information and description of data.
3. Data cleaning
    A. Handling of missing data
    B. Outlier handling
4. Imputation of data (if required)
    A. Handling null
    B. Handling unstructured data
5. Visualization
    A. Uni-variant
    B. Bi-variant
    C. Multi-variant
6. Data preparation before model building
7. Model building
8. Model evaluation
9. Prediction on test data

Understanding the Data

Understand the data in Leads.csv using shape, info and describe.

Data Reading & Data Types of Leads csv

```
#Read the data set of "Leads csv" in inpDay.
lead_Score=pd.read_csv("Leads.csv")
lead_Score.head()
```

Data Cleaning

- "Select" level represents null values for some categorical variables, as customers did not choose any option from the list.
- Check null values percentage
- Observe columns having > 40% missing values and its significance, if not require need to be dropped. Identify the columns which are not required in analysis, we can drop them to have better vision of data.
- Remaining column/variable with missing values need to be imputed by mode /Missing values.
- Columns with no use for modeling (Prospect ID, Lead Number) or only one category of response were dropped.
- Handle Outliers :Data Frame have many column/variable having Outliers. It should be handled properly in order to get better insight. Using boxplot to understand the data.

Data Imputation

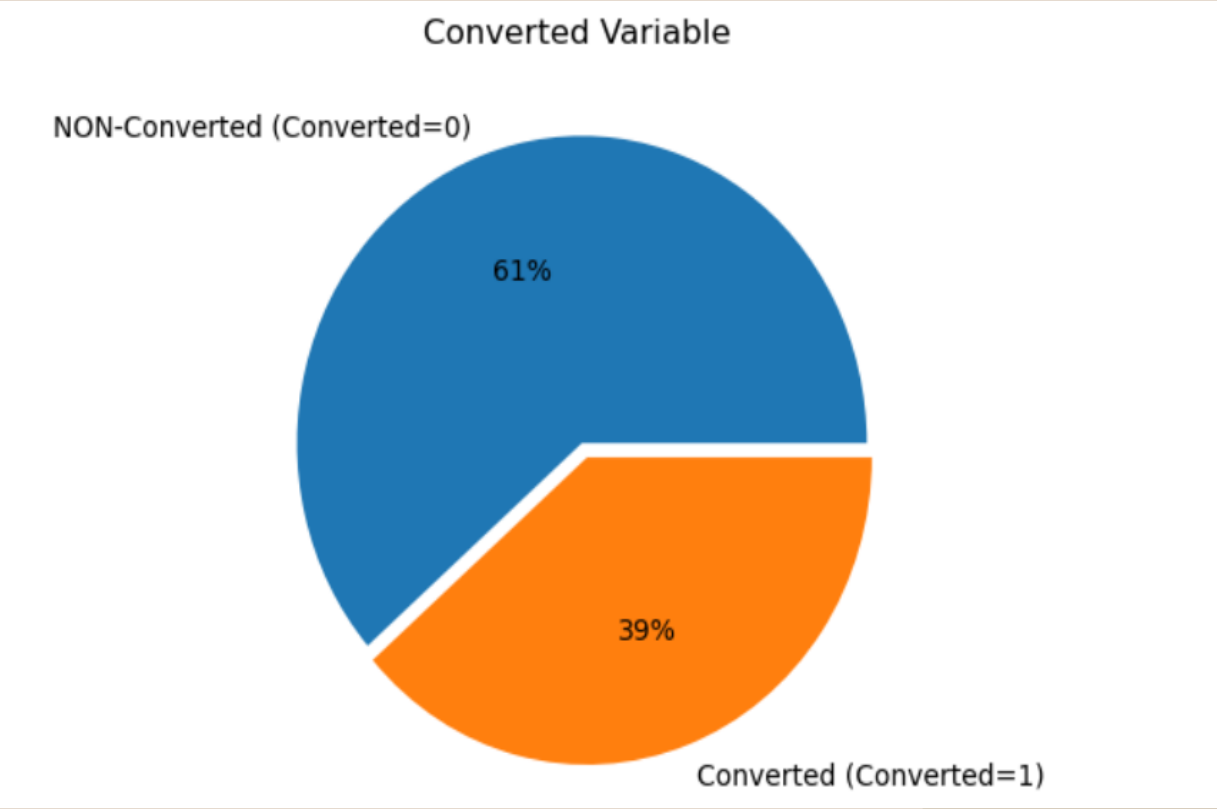
1. Handling Null

- It depends entirely on the type of missing value
- If the rows/column with the null is not going to affect the analysis or is in insignificant in number this can be drop from data frame.
- Fill the null values with mean median or mode as per the requirement or imputed as Missing/ unknown if its of MNAR type.

2. Handling Unstructured Data

- Involves preprocessing and analyzing data that does not have a predefined structure.
- Normalization: Convert text to lowercase or uppercase. Ex- Changed google to Google in Lead Source Column.
- Combined values to get more structured data.

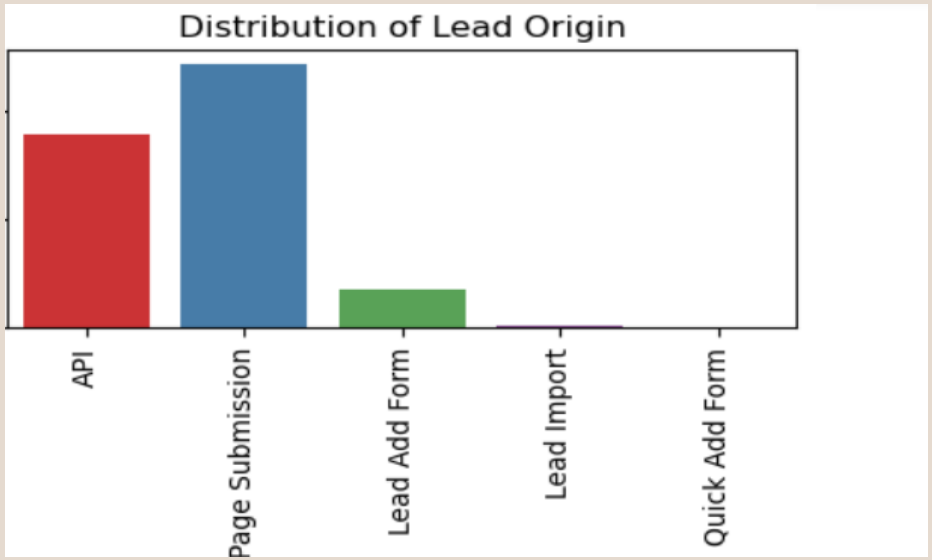
Data Imbalance

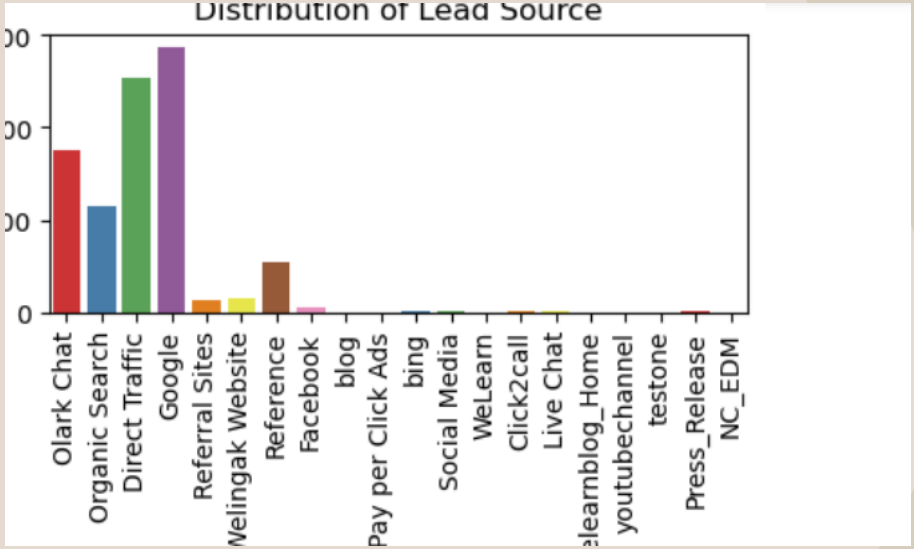


We can clearly see the that data set is highly imbalance. Conversion rate is of 39%, meaning only 39% of the people have converted to leads. While 61% of the people did not convert to leads.

- NON-Converted (Converted=0)
- Converted (Converted=1)

Univariant Analysis

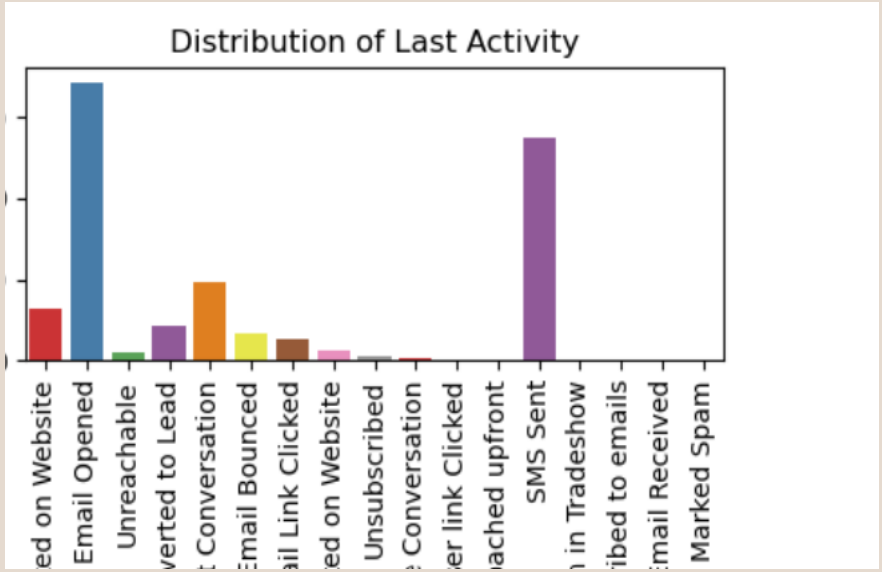


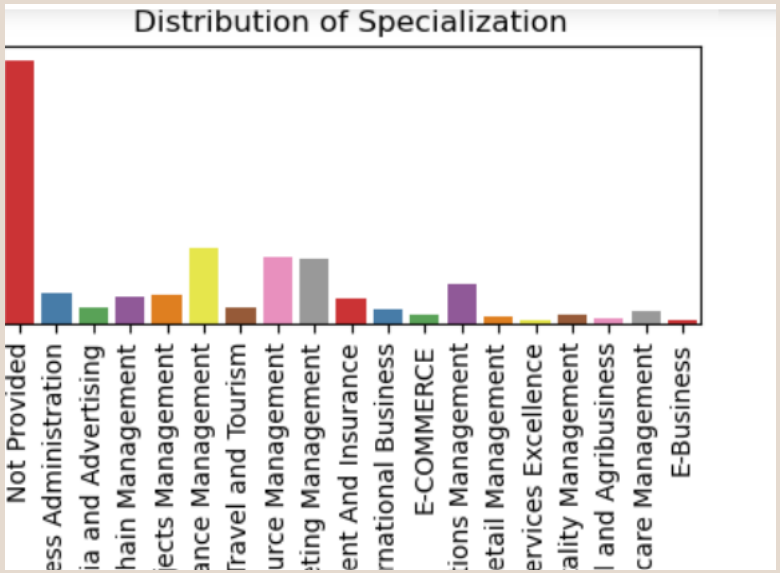


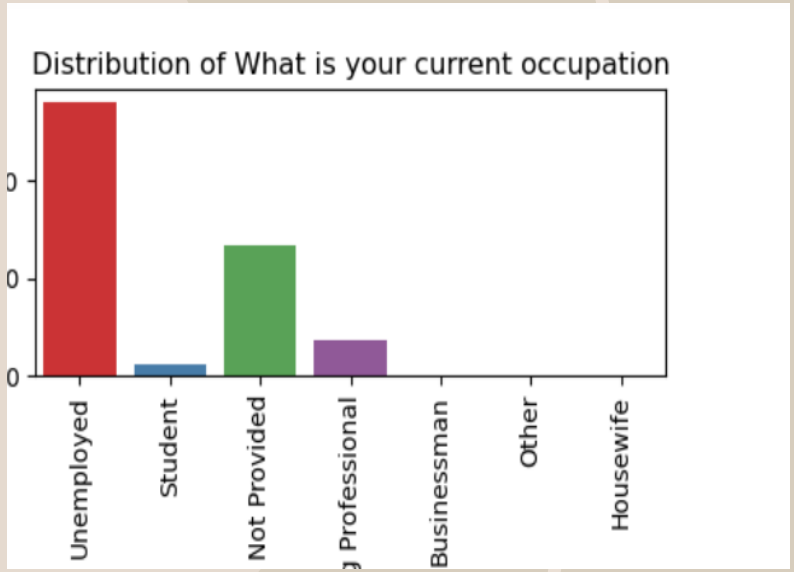
KEY INSIGHTS

1. Customer was identified to be a lead on the basis of their landing page submission.
2. The source of the lead majorly is from google search and direct traffic

Univariant Analysis



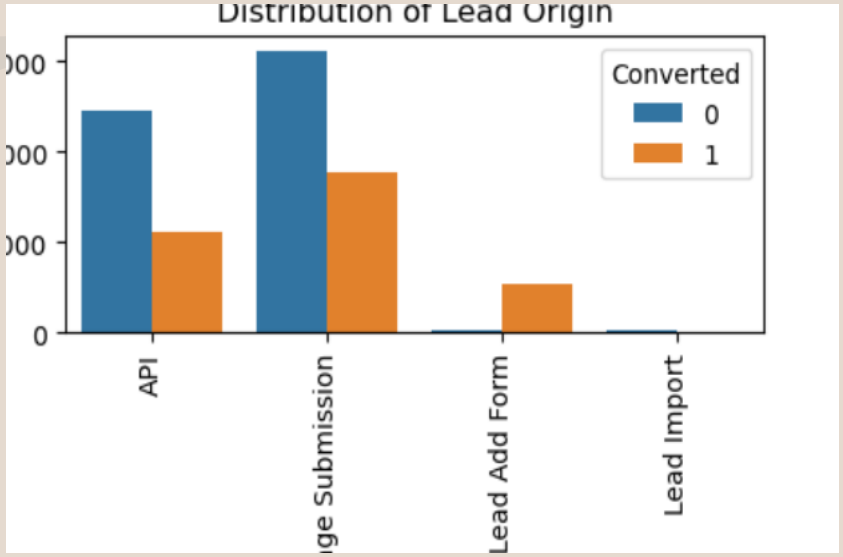


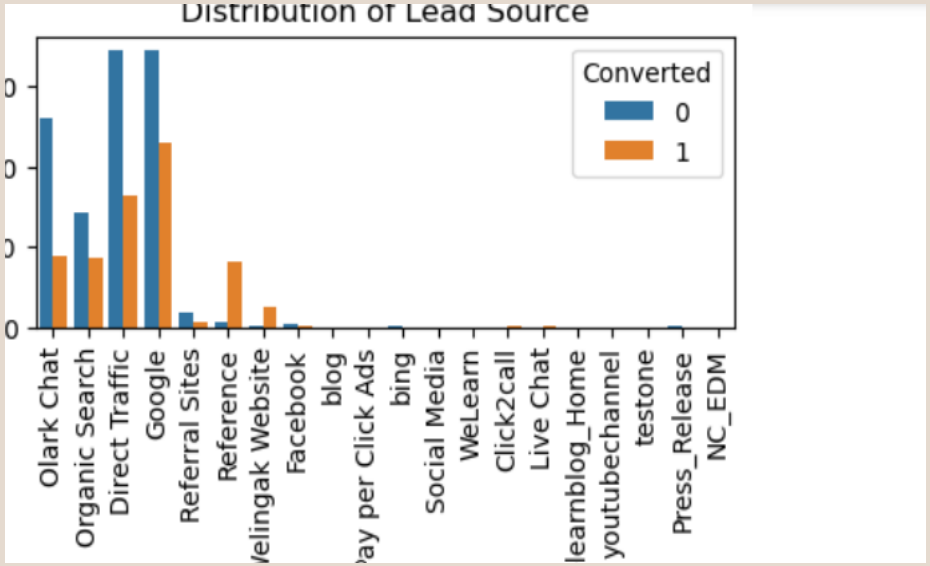


Key Insights

1. Last Activity of customers contribution in SMS Sent & Email Opened activities are on top.
2. Marketing Management, HR Management, Finance Management shows good contribution in Lead conversion.
3. Current occupation has more than 90% data for Unemployed.

Bivariant Analysis (Categorical)



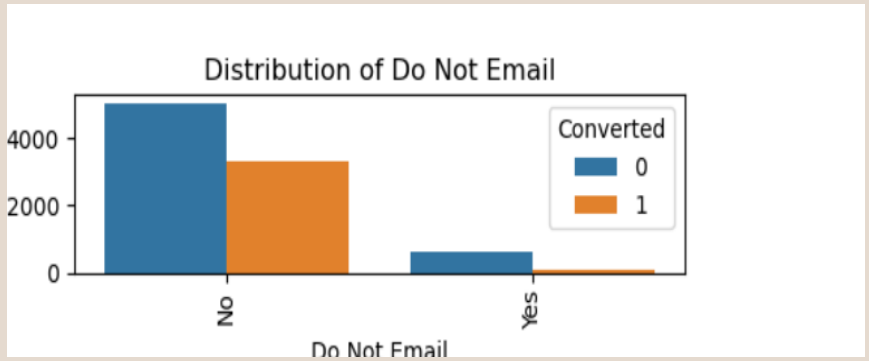


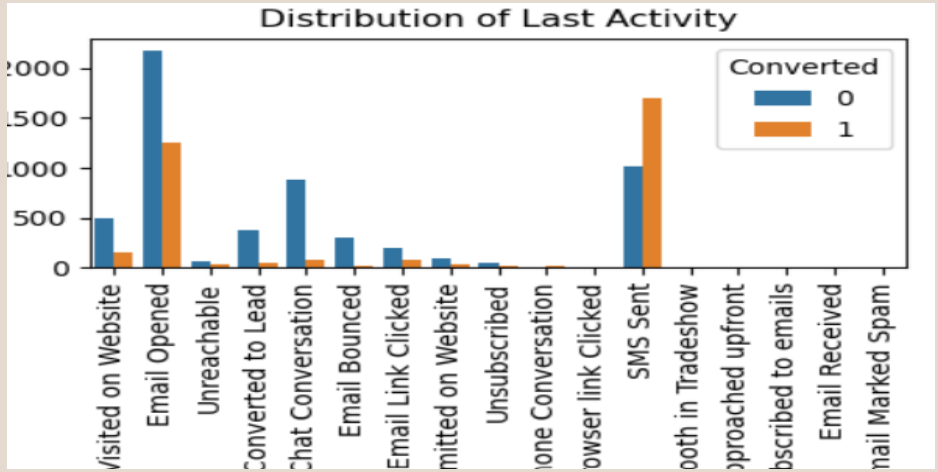
Key Insights

1. All leads originated from "Landing Page Submission" with a lead conversion rate (LCR) on lower side, the "API" identified as next lead factor but with lower conversion rate.

2. Google, Direct Traffic and Olark Chat are top lead source, however non conversion rate tends to be on higher side.

Bivariant Analysis



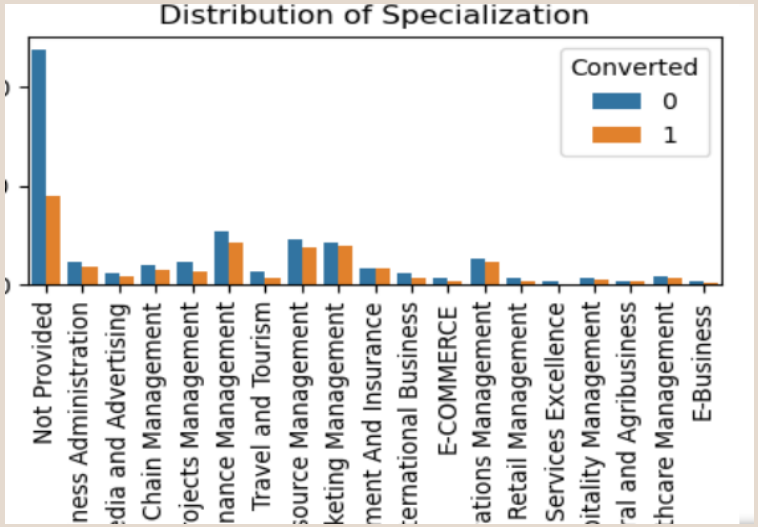


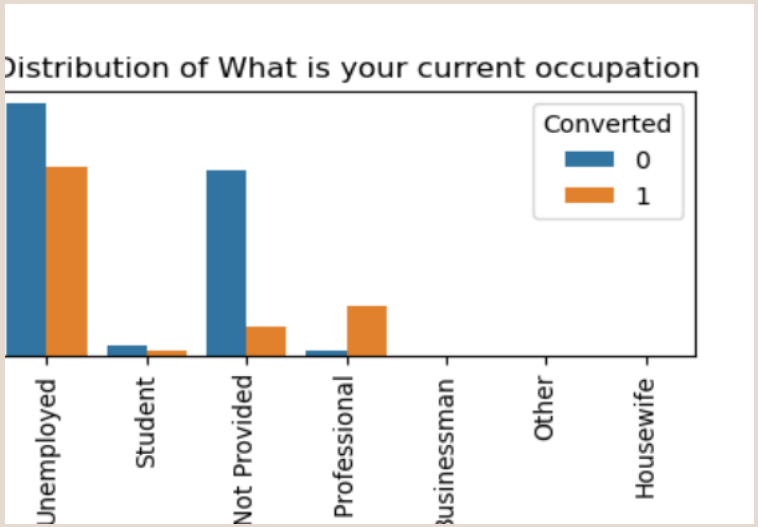
Keys Insights

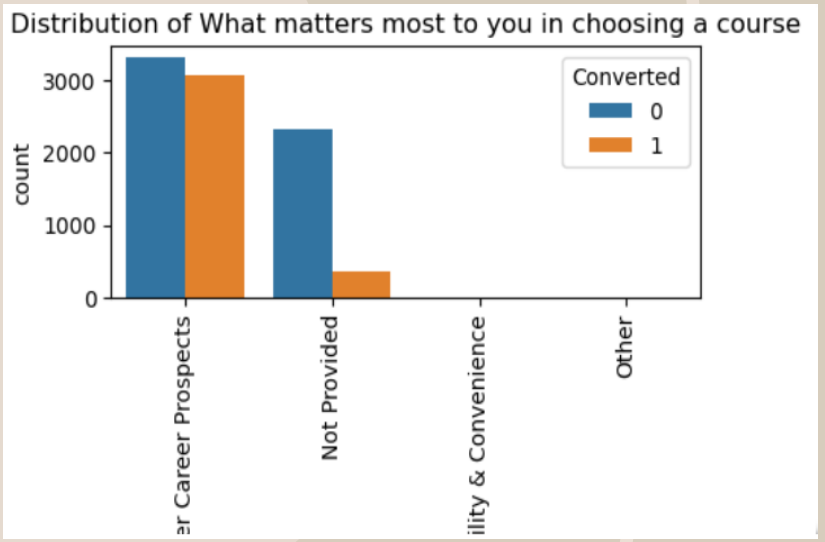
1. Though People have opted that they don't want to be emailed about the course, but still that leads to about 40% of them converting to leads.

2. Last activity of SMS sent has the highest conversion rate followed by Email opened.

Bivariant Analysis



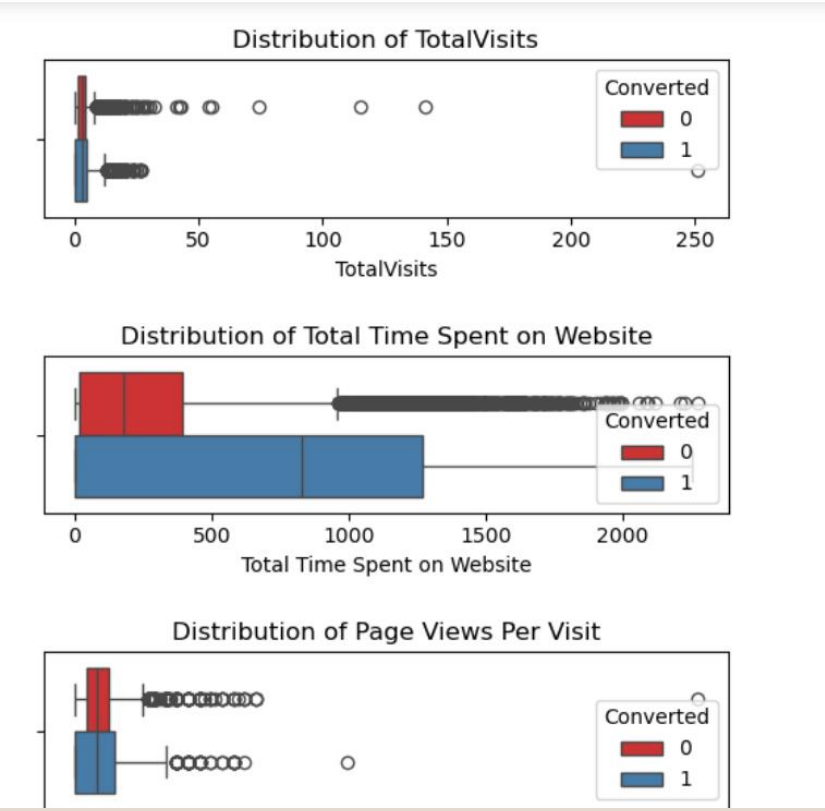




Key Insights

1. Marketing Management, HR Management, Finance Management shows good contribution in Leads conversion than other specialization.
2. The conversion rate is highest in Working professional, followed by unemployed.
3. Choosing Better Career prospect has highest conversion chances.

Bivariant Analysis (Numerical)

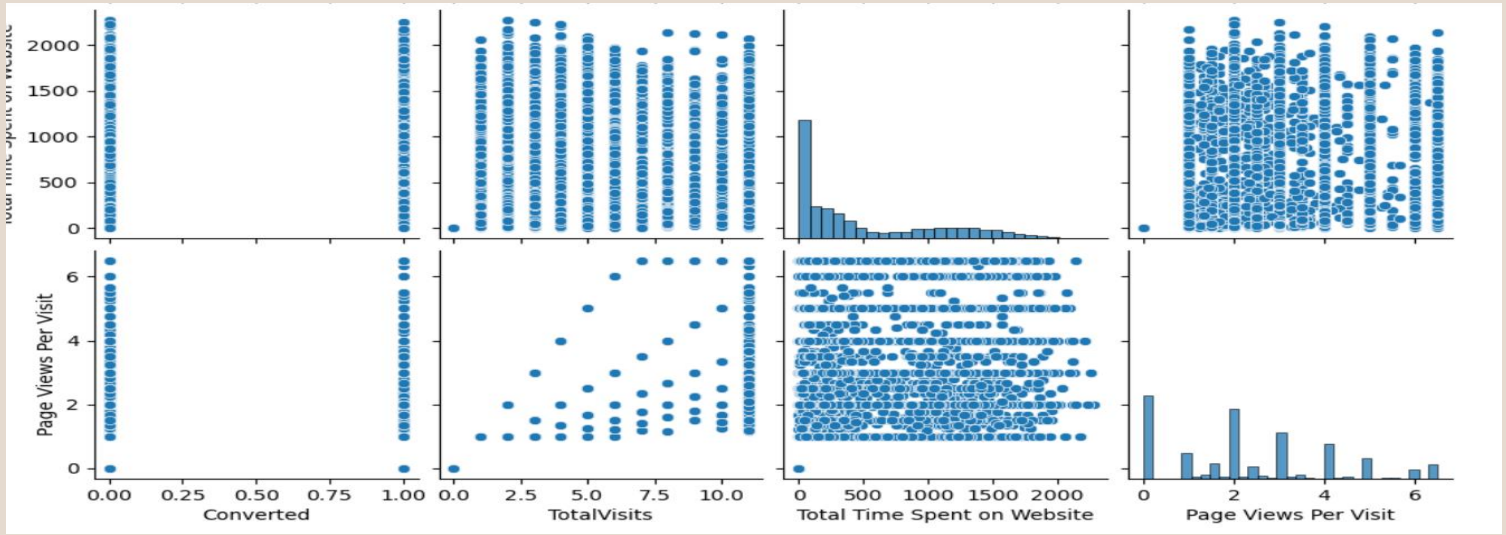


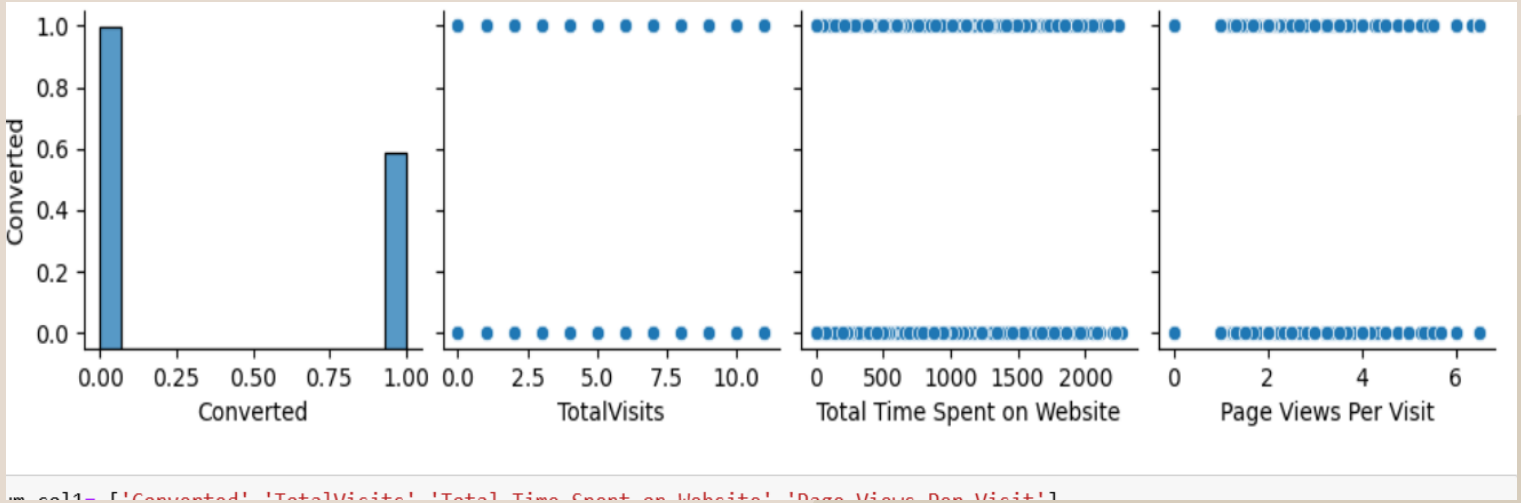
Key Insights:

1 Leads who spends more time on the Website have a higher chance of getting successfully converted than those who spends less time as seen in the box-plot

2 "TotalVisits","Page Views Per Visit": Both these variables contain outliers as can be seen in the boxplot So, These outliers needs to be treated for these variables

Bivariant Analysis (Numerical)

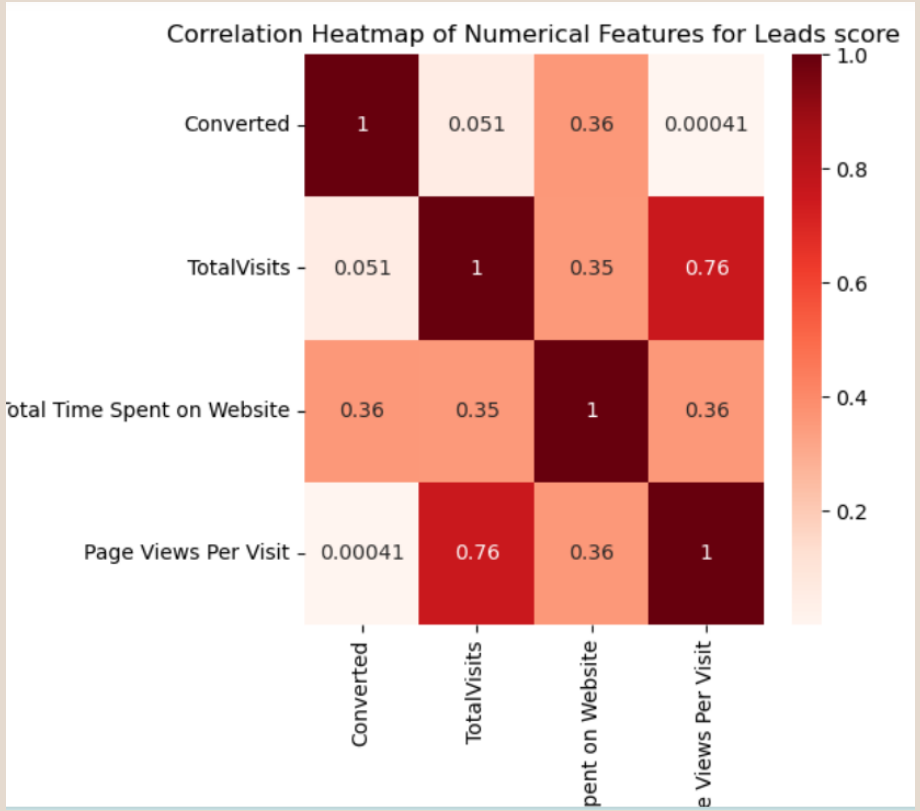




Key Insights:

There is positive correlation between Total Visit and Page Views per visit.

Multivariant Analysis



Key Insights:

1. Total Visits has strong correlation with page views per visit.

Data Preparation before Model building

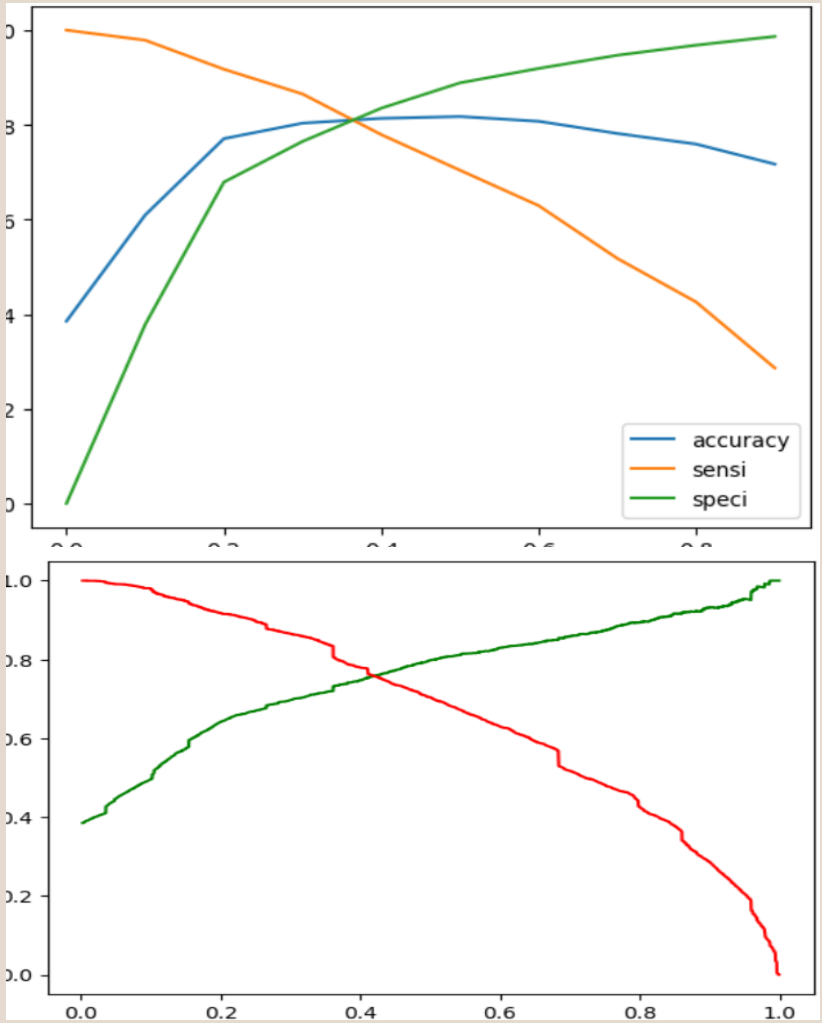
- Created dummy features (one-hot encoded) for categorical variables – lead origin, lead source,last activity, specialization, current_occupation
- Splitting train and test sets
- 70:30 % ratio was chosen for the split
- Feature scaling
  - ➢ Standardization method was used to scale the features
- Checking the correlations. Since there are a lot of variables it is difficult to find the corelation.

Model Building

- **Feature selection**
- The data set has large number of features , **feature selection using RFE TO SELECT IMPORTANT COLUMNS**
- Pre RFE – 83 columns and post RFE – 20 columns
- Implemented GLM MODEL
- Manual feature reduction process was used to build models by dropping variables with p – value greater than 0.05.
- Drop column with multicollinearity with VIFs greater than 5.
- Model 3 is stable model , and it was used for model evaluation which further will be used to make predictions.

Model Evaluation

Train data



With the current cut off as 0.5 we have around 82% accuracy, sensitivity of around 70% and specificity of around 89%.

After ROC curve cutoff is taken as 0.37.

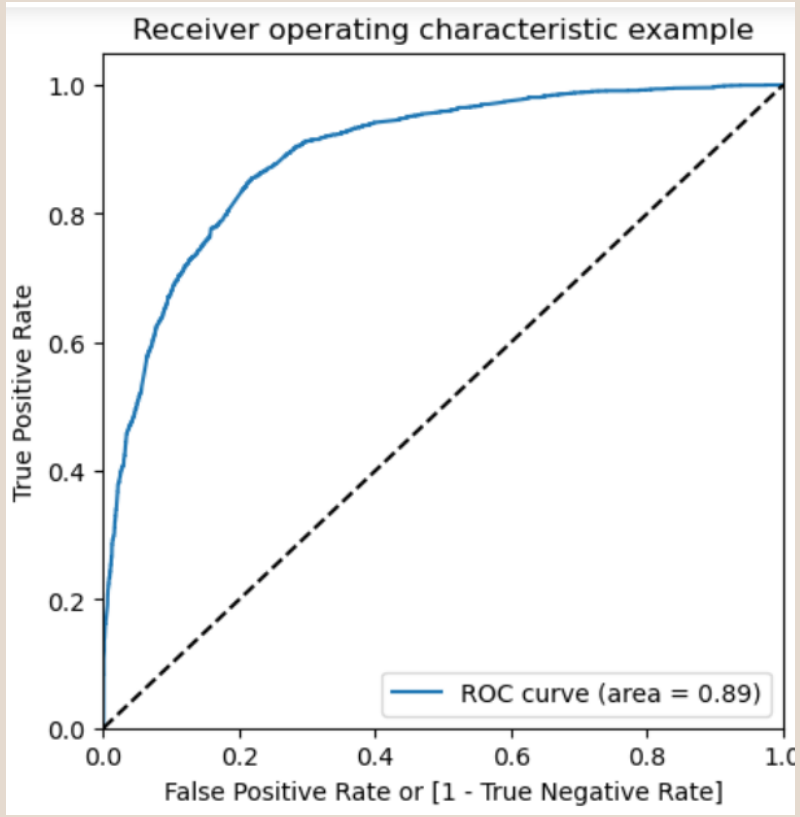
Overall Accuracy 81%
Confusion Matrix
array([[3203, 702],

[495, 1951]], dtype=int64)

Sensitivity 79.7%
Specificity 82%
Precision 73.5%
Recall 79.7%

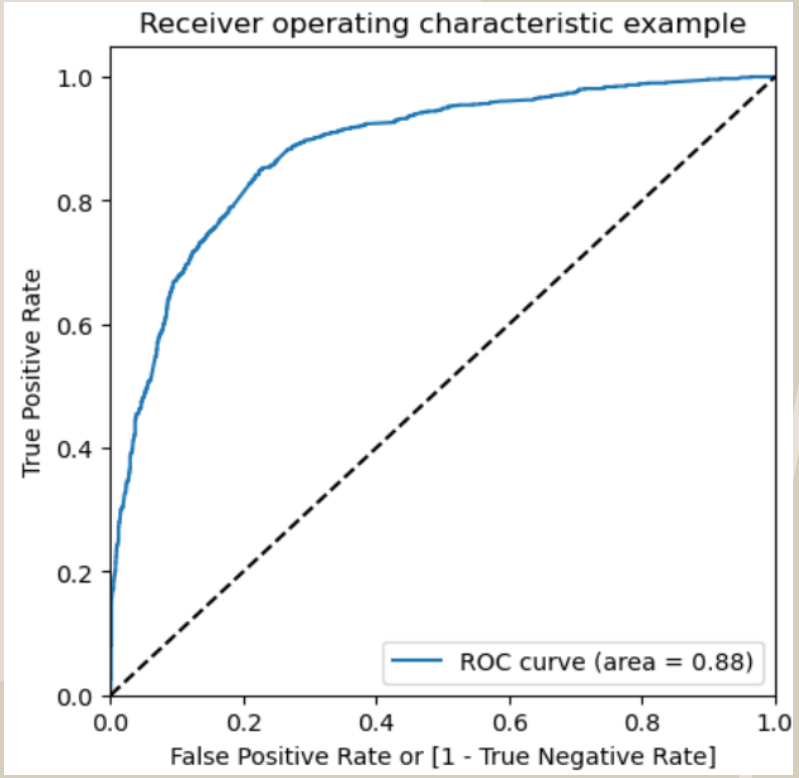
Model Evaluation

ROC curve – train data set



Area under ROC curve is 0.89 out of 1 which indicates a good predictive model.

ROC Curve – test data set



Area under ROC curve is 0.88 out of 1 which indicates a good predictive model.

Model Evaluation

Test data

CURRENT CUT OFF AS 0.37.

- **OVERALL ACCURACY** 80.9%
- **CONFUSION MATRIX**

ARRAY([[1424, 310],

[210, 779]], DTYPE=INT64)

- **SENSITIVITY** 78.7%
- **SPECIFICITY** 82%
- **PRECISION** 71.5%
- **RECALL** 78.7%

Conclusion

The variables that mattered the most in the potential buyers are :

1 The total time spend on the website.

2 When the lead source is:

- **A. Google**
- **B. Direct traffic**
- **C. Organic search**
- **D. Welingak website**
- **E. Referral sites**

3 When the last activity is

- **A. Olark chat conversation**

4 When the lead origin is lead add format.

5 Current occupation as a working professional students and unemployed candidates

6. Notable activity is email link clicked and email opened.

These variable x education can consider and achieve high chance to convert their all the potential buyers .

Recommendation

X Education need to increase its lead conversion for their growth and success.

To achieve this, we have developed a regression model that can help us identify the most significant factors that impact lead conversion.

We have determined the following features that have the highest positive coefficients, and these features should be given priority by their marketing and sales team to increase lead conversion:

- **OCCUPATION_WORKING PROFESSIONAL : 3.5743**
- **LEADORIGIN_LEAD ADD FORM : 2.3895**
- **LEADSOURCE_WELINGAK WEBSITE: 2.0246**

Recommendation

- **OCCUPATION_STUDENT: 1.2187**
- **TOTAL TIME SPENT ON WEBSITE :1.1349**
- **OCCUPATION_UNEMPLOYED: 1.1313**

We have also identified features with negative coefficients that may indicate potential areas for improvement. These include:

- **DONOTEMAIL_YES: -1.8723**
- **NOTABLEACTIVITY_EMAIL LINK CLICKED: -1.7025**
- **NOTABLEACTIVITY_EMAIL OPENED: -1.3407**
- **LEADSOURCE_DIRECT TRAFFIC:-1.6361**

Recommendation

- **LEADSOURCE_REFERRAL SITES:-1.4946**
- **LASTACTIVITY_OLARK CHAT CONVERSATION:-1.1587**
- **NOTABLEACTIVITY_PAGE VISITED ON WEBSITE:-1.4730**

Hence, focus on features with positive coefficients for targeted marketing strategies and analyzing negative coefficients for improvement will definitely drive the conversion rate higher

Thank you