LEAD SCORE ASSIGNMENT

upGrad & IIITB
Data Science
Program
May 2023

Sneha Mendon
Jeevitha
Naga Sri Muvva

TABLE OF CONTENTS

1. Problem Statement

- Create a logistic regression model for X Education to optimize lead conversion. By assigning lead scores (0 to 100) using historical data attributes, the model aims to pinpoint hot leads with a greater probability of conversion. This initiative seeks to significantly boost the lead conversion rate, aligning with the company's ambitious target of 80%.

2. Data Cleaning & PreProcessing

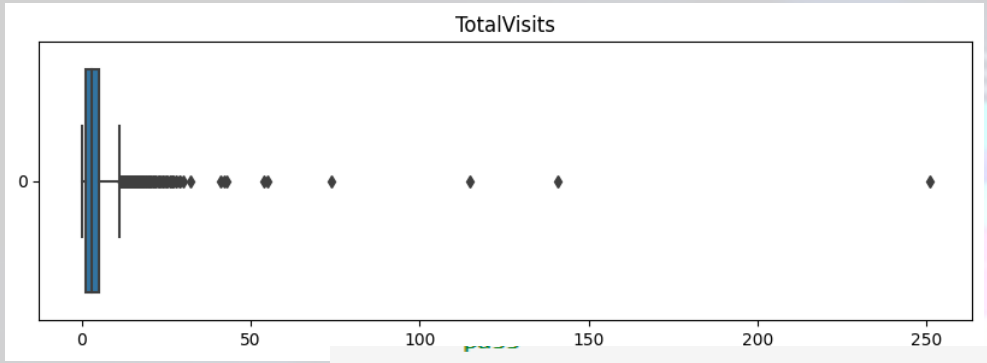
- Data Cleaning
  - Columns “Specialization”,” How did you hear about X Education”,” Lead Profile”,”City” have value ‘select’ which implies not selected. We have replaced them with NaN.
  - Grouped Low frequency values in categorical variables to “Others”
  - Imputed NaN which are >30% to “NA”

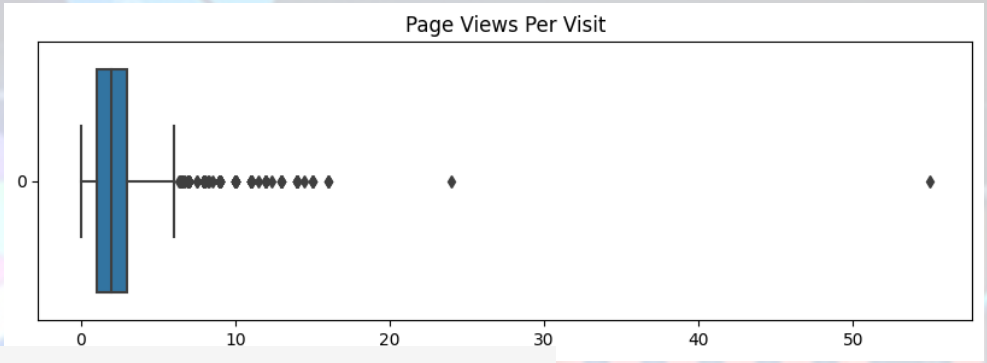
2. Data Cleaning & PreProcessing

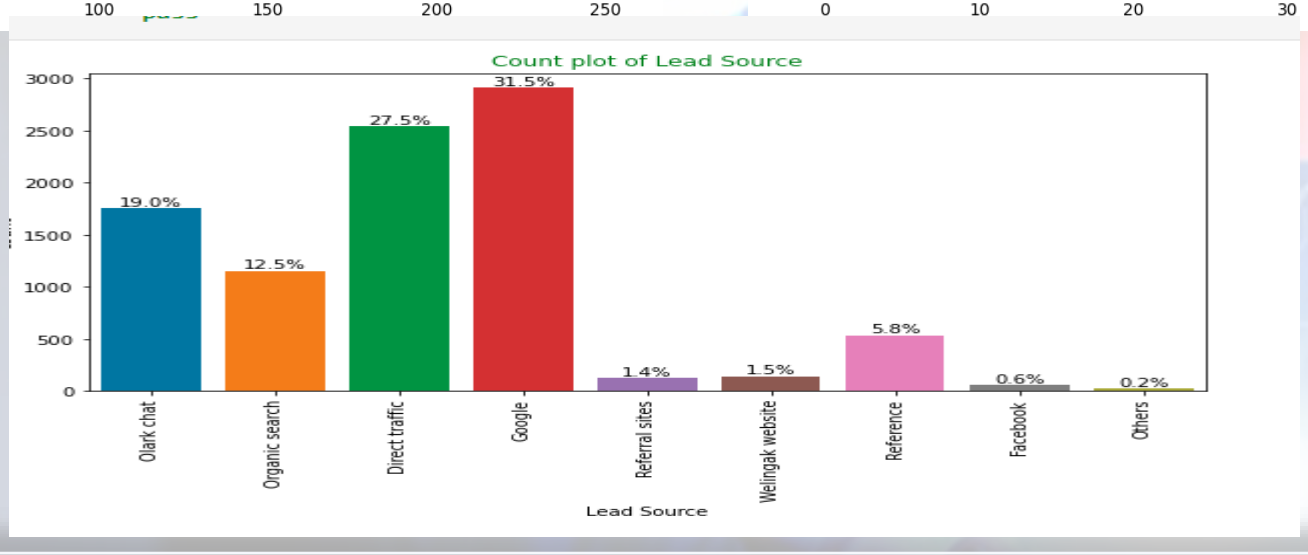
- Data Cleaning – Missing Values
  - ✓ Variables having more than 40% of missing data are removed
  - ✓ Variables having <30% missing values are imputed with Mode(all are categorical columns)
- Outlier Detection
  - ✓ Columns like TotalVisists, Page Views Per Visit have unlikely extreme values which are impossible in reality, they can be replaced with quartile calculations.

Exploratory Data Analysis

Uni Variate Analysis



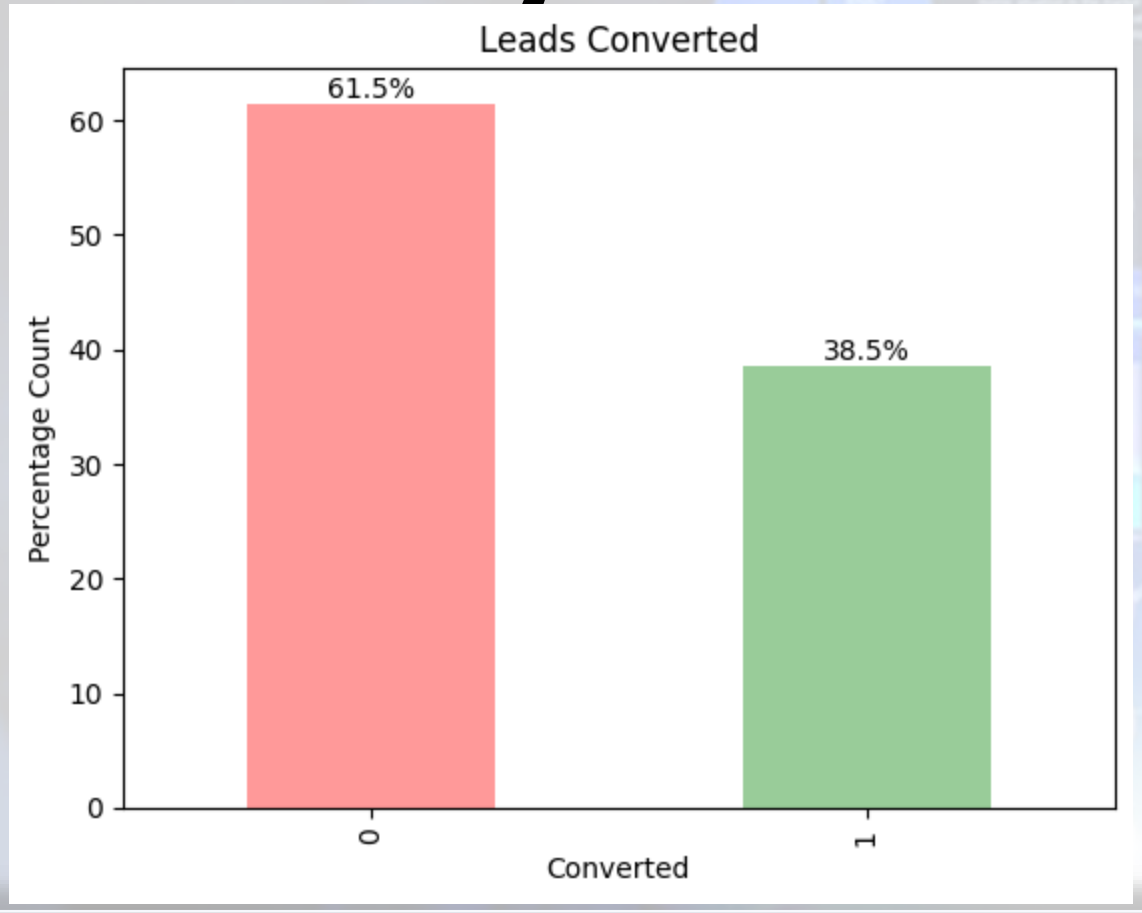




Inferences:

- TotalVisits, Page Views Per Visit have potential outliers and can be replaced.
- Lead source shows that we have more sources from google.

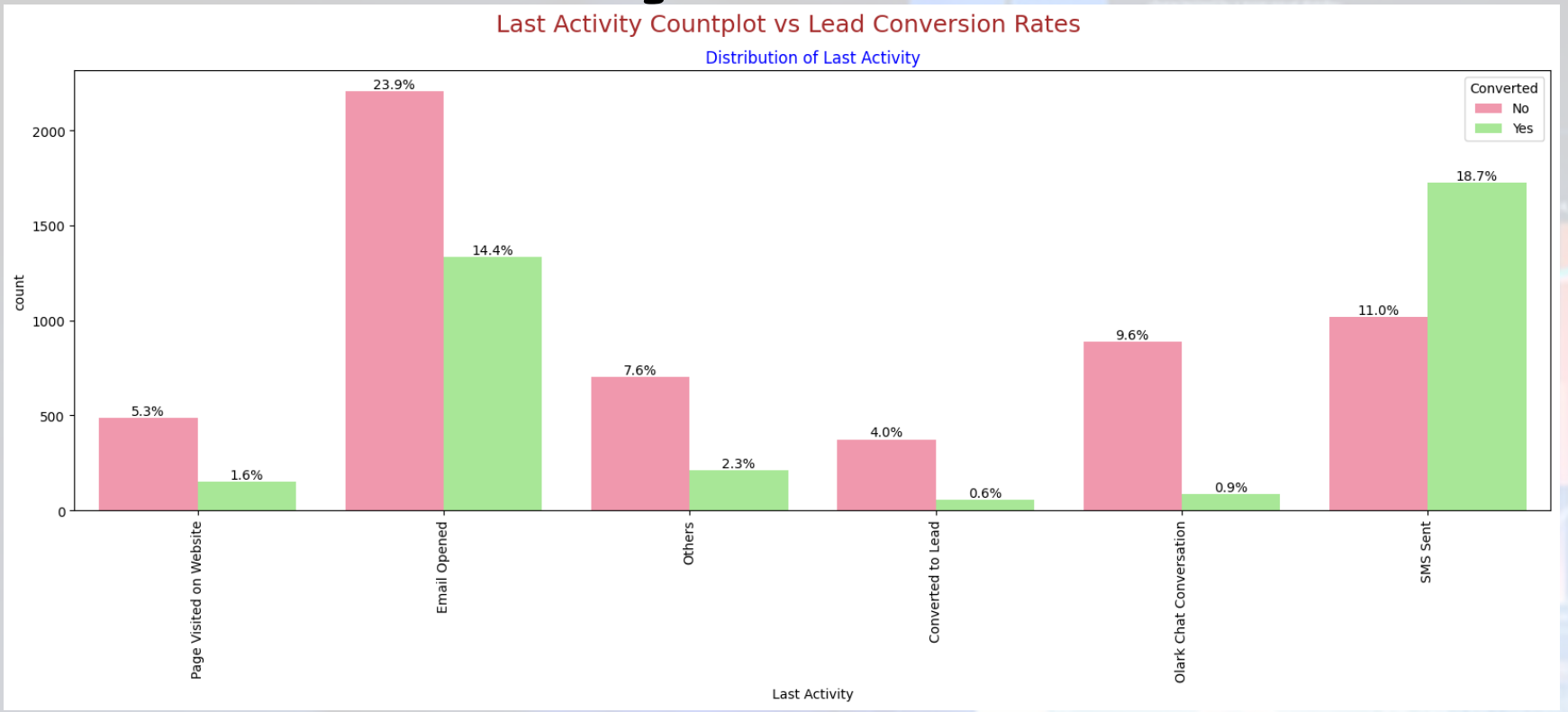
Uni Variate Analysis



Inferences:

- Data is imbalanced while analyzing target variable.
- Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)
- While 61.5% of the people didn't convert to leads. (Majority)

Bi Variate Analysis



Inferences:

- Last Activity SMS sent has a significant positive affect on target variable

4. Data Preparation

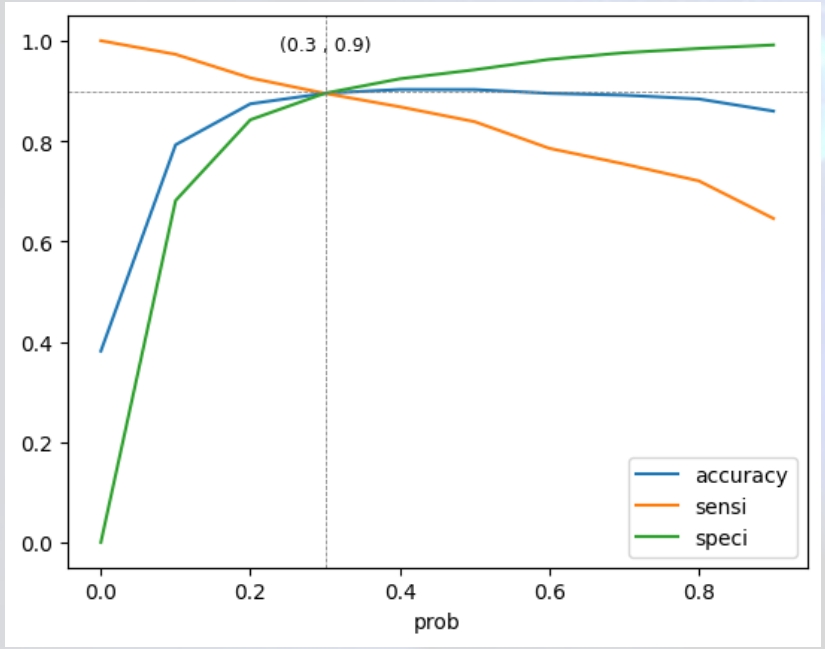
- Binary level categorical columns were already mapped to 1 / 0 in previous steps
- Created dummy features (one-hot encoded) for categorical variables – 'Lead Origin','Lead Source','Last Activity','Specialization','What is your current occupation','Tags','City','Last Notable Activity'
- Splitting Train & Test Sets ○ 70:30 % ratio was chosen for the split
- Feature scaling ○ Standardization method was used to scale the features
- Checking the correlations ○ Predictor variables which were highly correlated with each other were dropped (Last Notable Activity_SMS Sent and Lead Origin Lead Add Form)

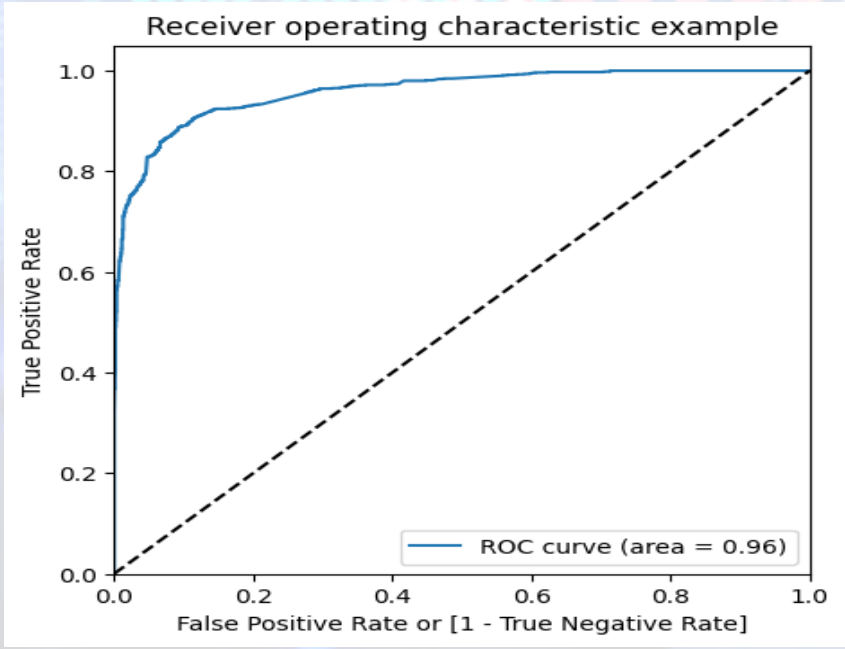
5. Model Building

- The data set has lots of dimension and large number of features.
- This will reduce model performance and might take high computation time.
- Hence it is important to perform Recursive Feature Elimination (RFE) and to select only the important columns.
- Then we can manually fine tune the model.
- RFE outcome ○ Pre RFE – 44 columns & Post RFE – 15 column
- Model 5 looks stable after four iteration with: ○significant p-values within the threshold (p-values < 0.05) and ○No sign of multicollinearity with VIFs less than 5
- Hence, logm5 will be our final model, and we will use it for Model Evaluation which further will be used to make predictions

6. Model Evaluation

- Confusion Matrix & Evaluation Metrics with 0.3 as cutoff Confusion Matrix & Evaluation Metrics with 0.39 as cutoff It was decided to go ahead with 0.3 as cutoff after checking evaluation metrics coming from both plot.
- Area under ROC curve is 0.88 out of 1 which indicates a good predictive model.
- Using a cut-off value of 0.3, the model achieved a sensitivity of 84.15% in the train set and 88.86% in the test set.
- The model also achieved the test accuracy of 89.75% which is more than expected(80%).





6. Recommendations

- As per the problem statement, increasing lead conversion is crucial for the growth and success of X Education. 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 in our marketing and sales efforts to increase lead conversion.
    - Tags_Closed by Horizzon 10.405217
    - Tags_Will revert after reading the email 7.437032
    - Tags_Others 3.484836
    - Tags_NA 3.041428
    - Last Activity_SMS Sent 1.908150
    - Lead Source_Others 1.524952
    - What is your current occupation_Working Professional 1.011007
    - Total Time Spent on Website 0.921281
    - Lead Origin_Landing Page Submission -1.053471
    - Last Notable Activity_Modified -1.299453
    - Lead Origin_Lead Import -2.271064

THANK YOU