LEAD SCORING CASE STUDY

by

Narendran Balasundaram
Narra Siva Sai Kumar
Avinash

PROBLEM STATEMENT:

- X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

GOAL:

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future

STEPS PERFORMED IN OUR ANALYSIS:

I) READING AND UNDERSTANDING DATA

II) EXPLORATORY DATA ANALYSIS

III) DATA PREPARATION

IV) TEST-TRAIN SPLIT

V) FEATURE RESCALING

VI) MODEL BUILDING

VII) MODEL EVALUATION

I) READING AND UNDERSTANDING

DATA: •

- We have Imported the libraries that are required for our analysis.

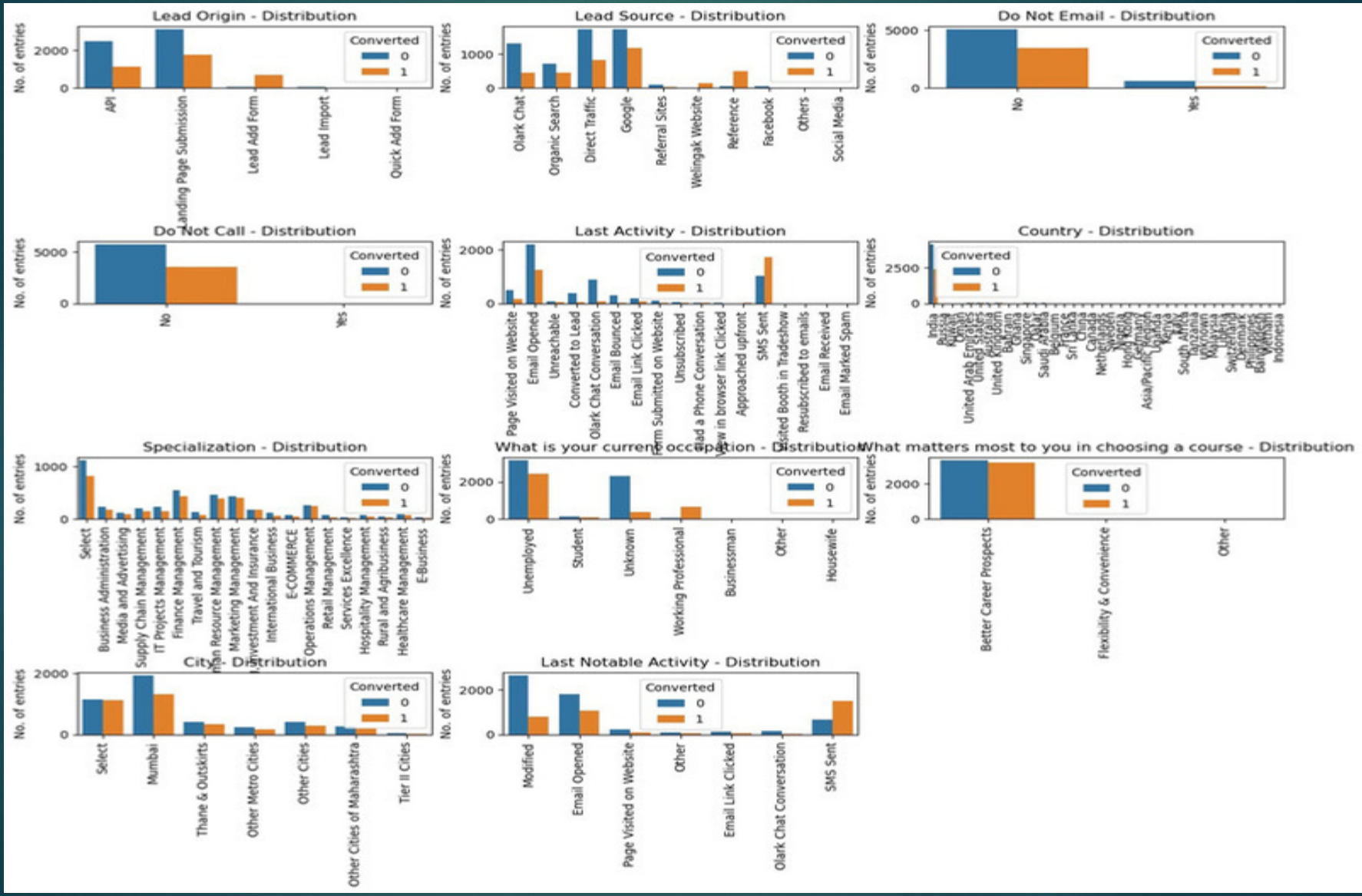
  Initially we have performed simple data checks on the given raw data.

II) EXPLORATORY DATA ANALYSIS: •

- We have performed data Cleaning and outlier treatments on raw data.
- We have performed Univariate and Bivariate Analysis on Categorical and Numerical Columns.

On upcoming slides we discuss about our categorical and numerical analysis.

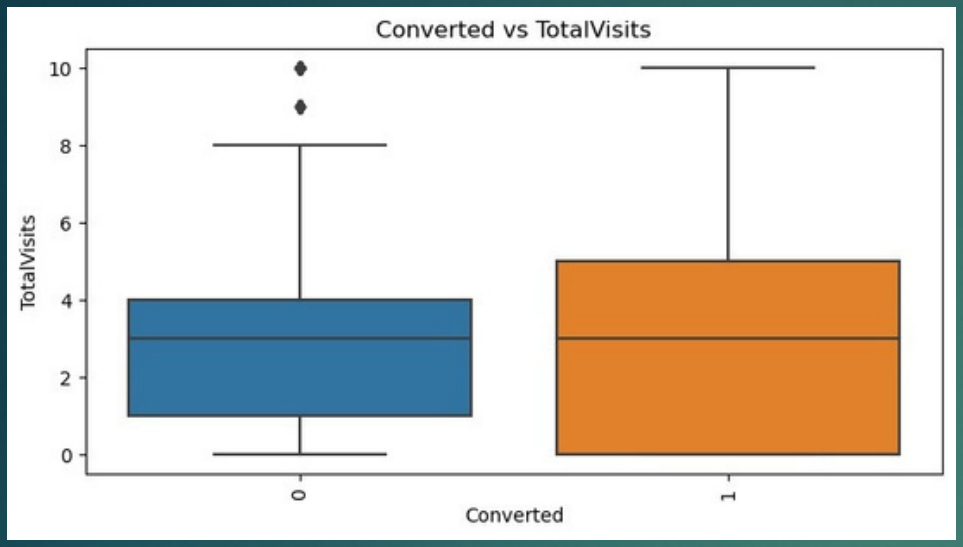
CATEGORICAL ANALYSIS:

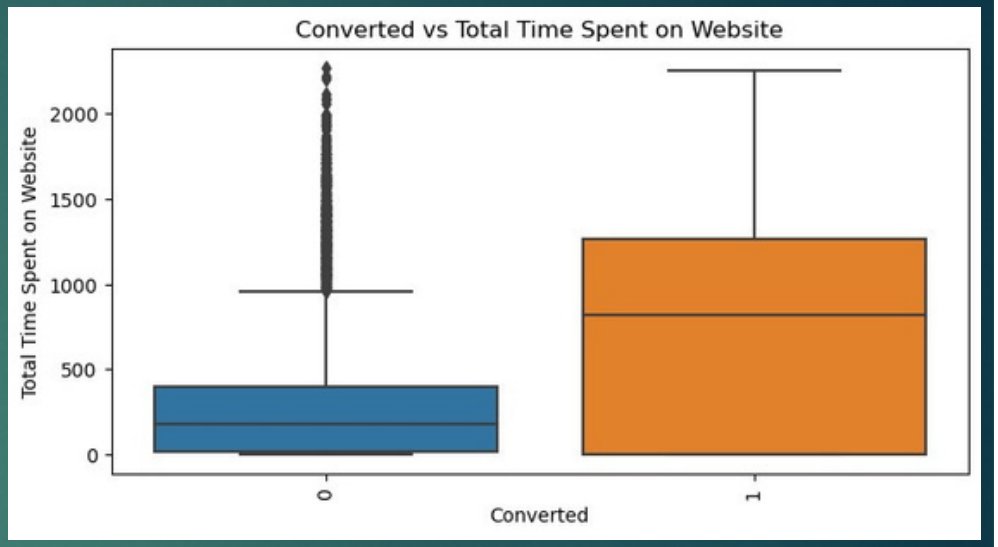


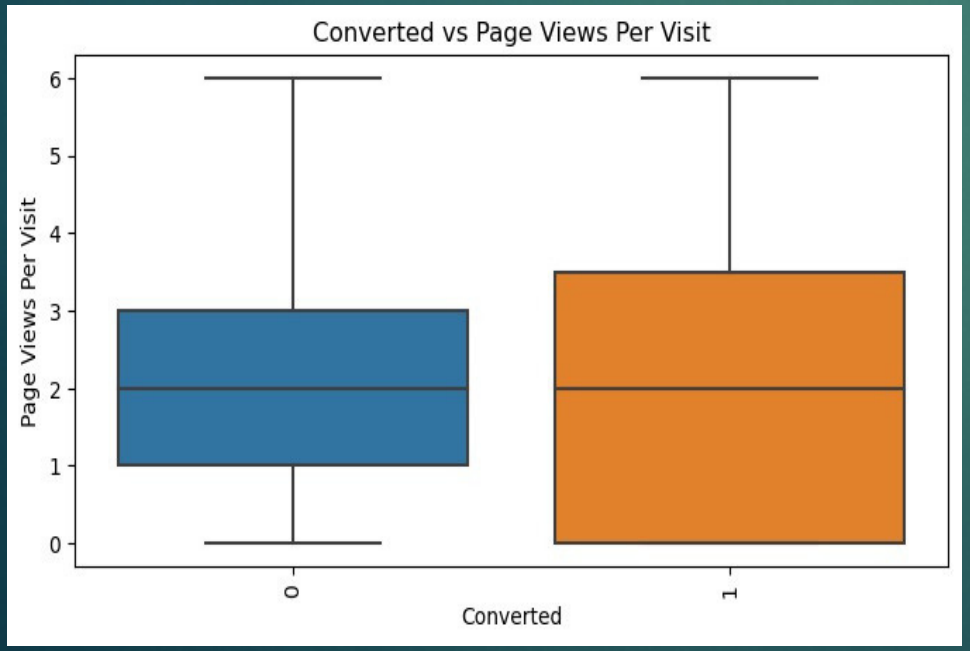
Observations on Cat. columns from our analysis:

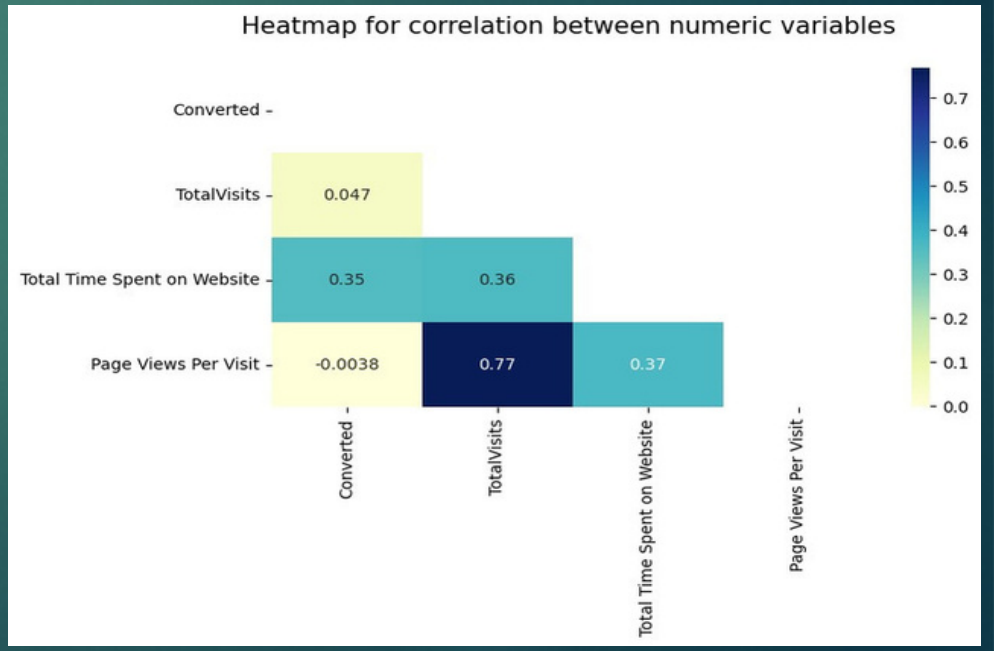
- Do Not Call: From the above plot we are able to see imbalance therefore, we can drop the column
- Last Activity: Email and SMS has high frequency based on their last Activity
- Country: Its mostly India, no inference can be drawn from this parameter.
- Specialization: We cannot conclude/impute this as it contains null values in large amount and other disciplines are distributed some what equal.
- What matters most to you in choosing a course: Its mostly Career improvement, no inference can be drawn from this parameter.
- City: Mumbai holds the first position and however we don't have full info as more than half of our values are null(i.e'Select')
- Last Notable Activity: Maximum leads are generated having last activity as Modified and Email opened .

NUMERICAL COLUMN : BIVARIATE and HEATMAP









Observations on Num. columns from our analysis:

- From Converted vs Total visits and Converted vs Pageviews per visit plots, we cannot conclude anything since Median for converted and not converted leads is almost same.
- From Converted vs Total time Spent on Website, we can say that leads who spends most time on website are more likely to convert , thus website should be made more enagagingto increase conversion rate.
- From HeatMap, we observe that Page views per visit and Total visits are highly correlated

SUMMARY ON OUR EDA:

- Data cleaning and data handling has beenddone on our raw data.
- We have performed categorical analysis and dropped columns which are not required or irrelevant for our analysis.
- Also we have performed analysis on numerical variables, bivariate analysis on Num. variables
- Now, all data labels are in good shape , we will proceed to our next step-Data Preparation

III) DATA PREPARATION:

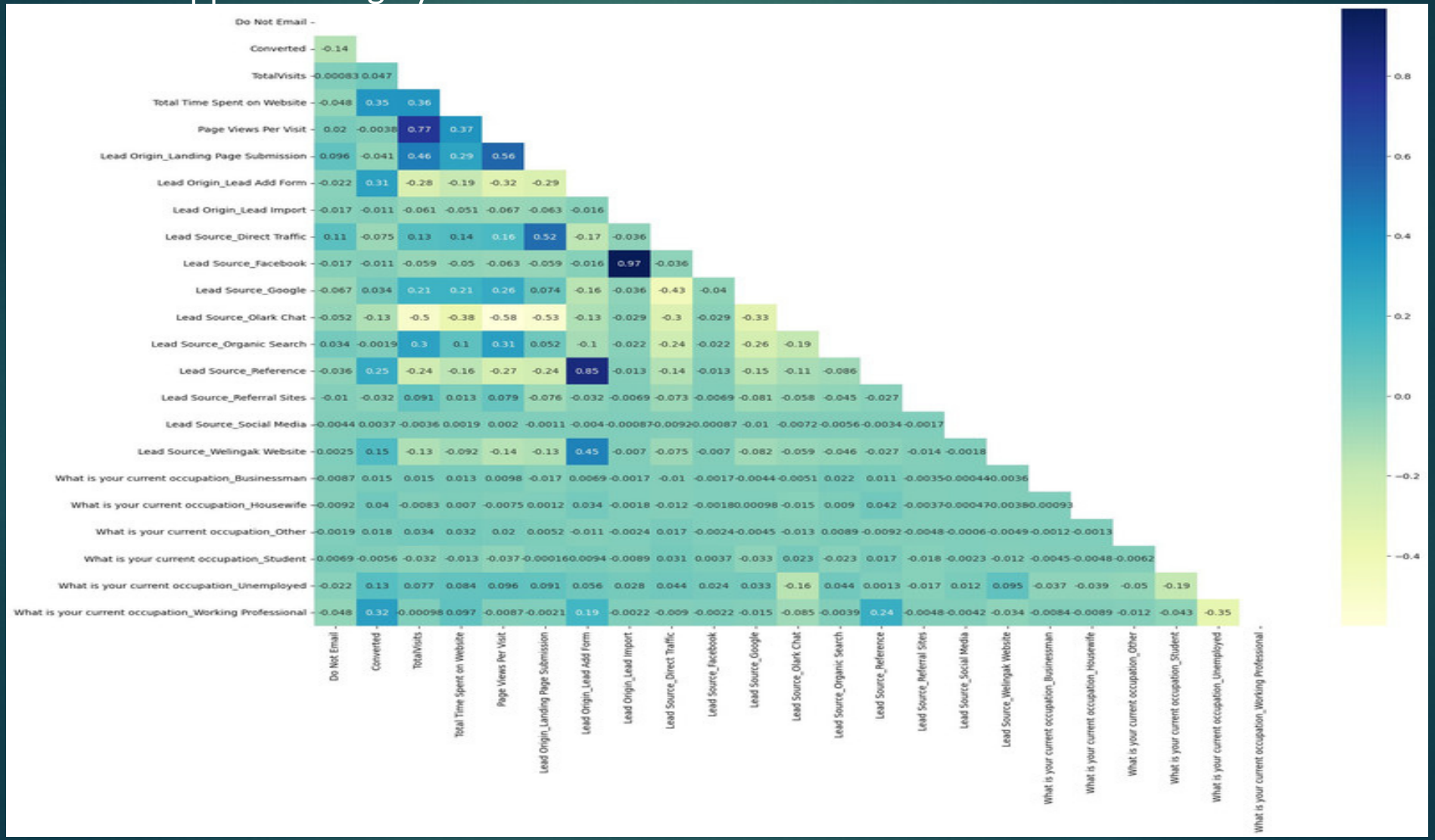
- In our Dataframe"Do Not Email" column we have Yes or no values. So we have converted them to 0/1 respectively.
- We created dummy variables for the categorical columns. Removed all the repeated and redundant variables

IV) TEST-TRAIN SPLIT

- We have divided the dataset into test and train sections with a proportion of 70- 30% values.

V) FEATURE RESCALING: •

With the dummy variables we have plotted the heatmap after scaling them and dropped the highly correlated variables.



VI) MODEL BUILDING: •

Using the Recursive Feature Elimination (RFE) approach, we have built the models.

- And selected the model which is efficient from our approach.
- We are proceeding our finalized model for Model Evaluations.

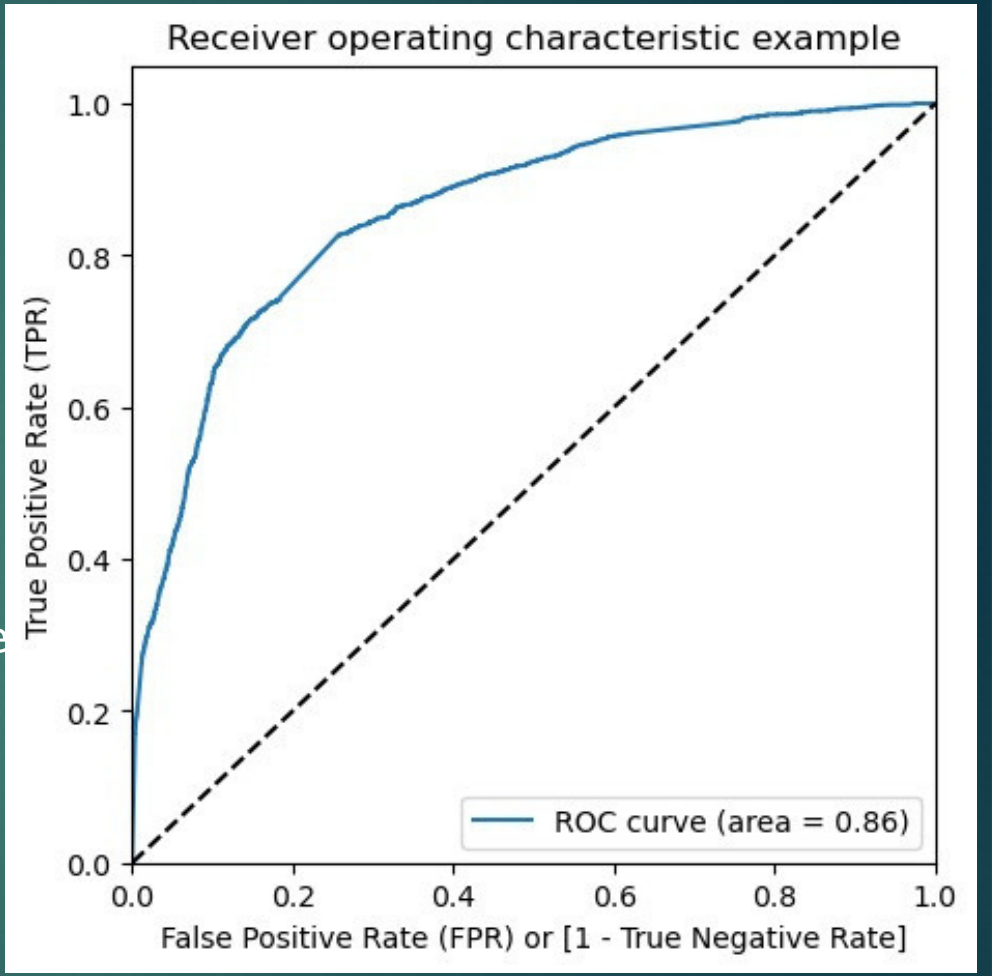
VII) MODEL EVALUATIONS:

•Predicted test and train models.

•Compared the metrics of above to conclude.

PLOTTING ROC CURVE:

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- The ROC Curve should be a value close to 1. We are getting a good value of 0.86 indicating a good predictive model



COMPARISION OF METRICS :

Contents	TRAIN DATA SET (approx.)	TEST DATA SET (approx.)
ACCURACY	77.05	77.52
SENSITIVITY	82.89	83.01
SPECIFICITY	73.49	74.13

SUMMARY:

- Upon checking both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.

- Accuracy, Sensitivity and Specificity values of test set are around 77%, 83% and 74% which are approximately near to the respective values calculated using trained set.
- Lead score is calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%

- Therefore, we can consider this model as a good one.

- Top 3 Important features responsible/helps for good conversion rate are:

  * Lead Origin_LeadAdd Form
  * What is your current occupation_WorkingProfessional
  * Total Time Spent on Website