Lead Scoring Case Study

Submitted By:
Anish Gupta
Ninad Milind Kalambe
Aachal Rajesh Deshmukh

Problem Statement

- An education company named X sells online courses to industry professionals. They have a process of online form filling on their website.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead. Once acquired, employees from the sales team start making calls, writing emails, etc. and some of the leads get converted while most do not.
- The typical lead conversion rate at X education is around 30% which is poor.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.

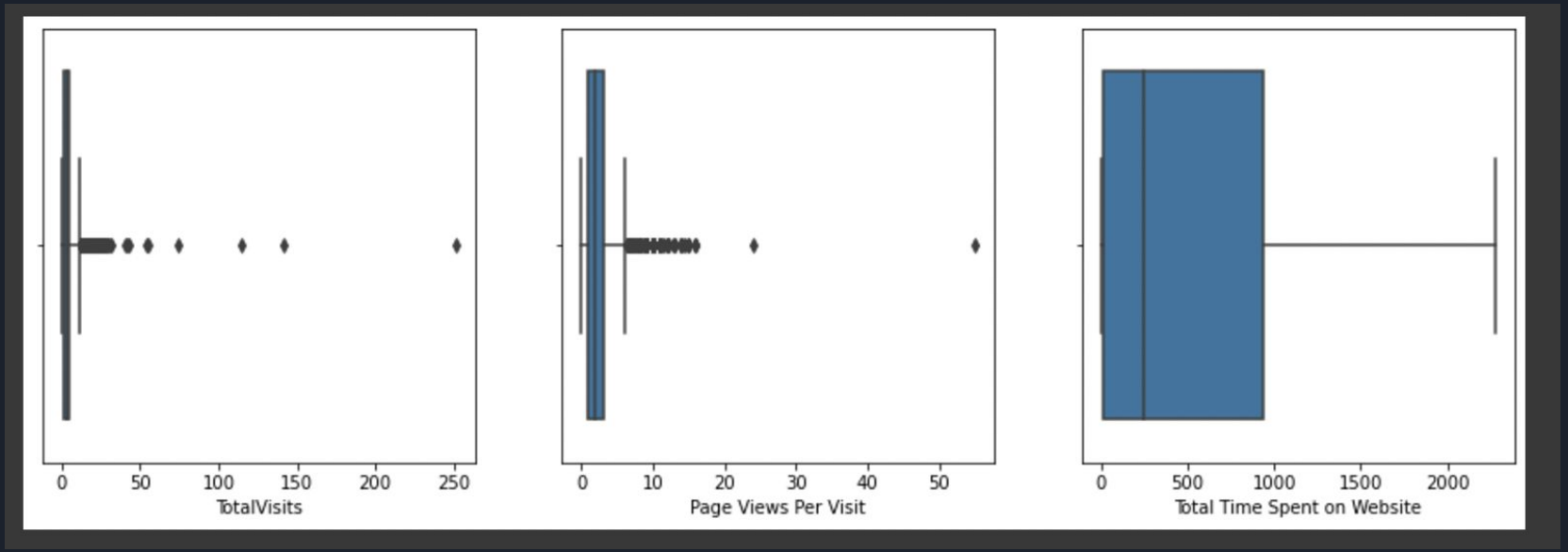
Business Objective

- X Education wants 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 a 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.
- The CEO wants to achieve the target lead conversion rate to be around 80%.

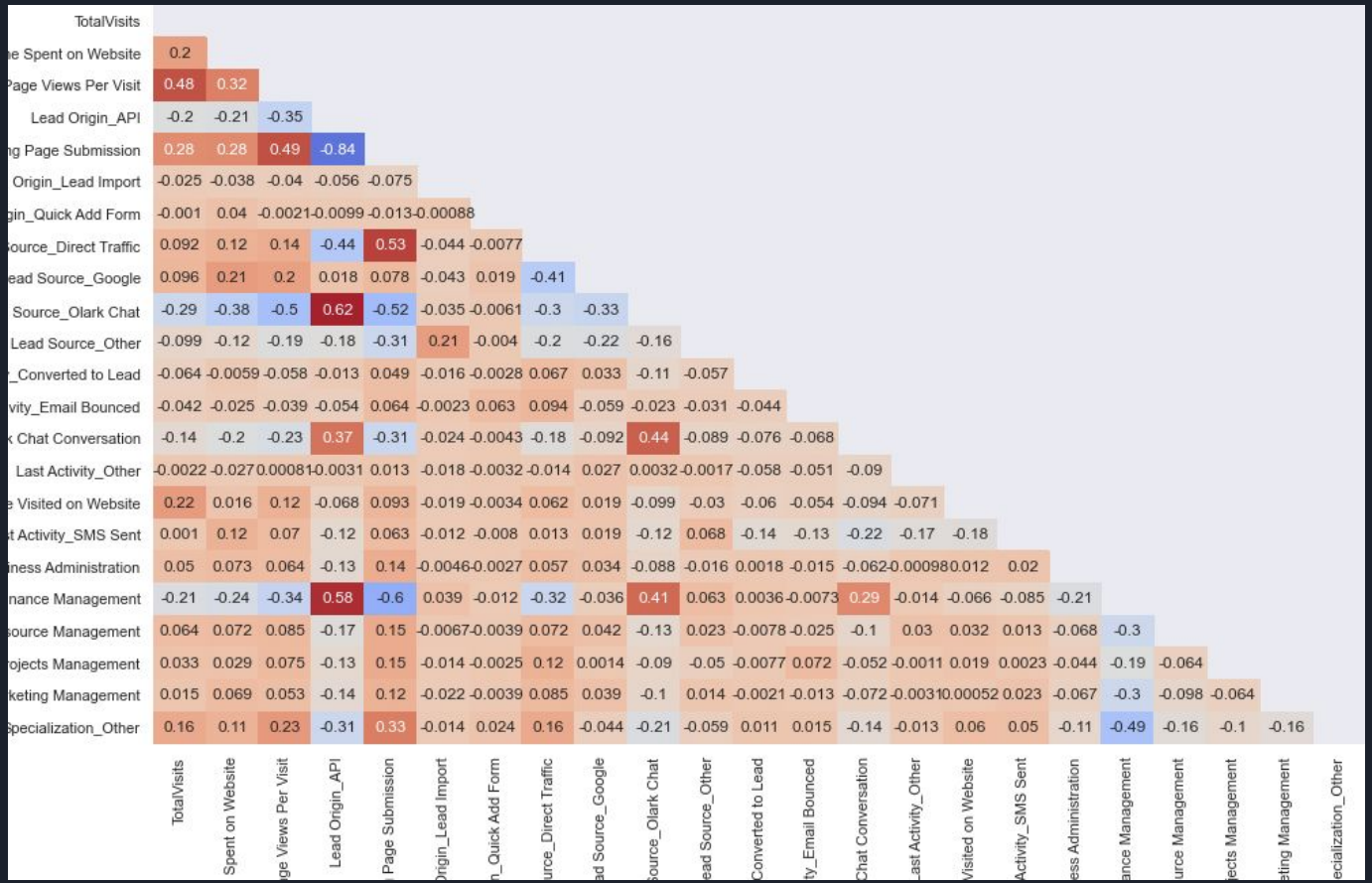
Analysis Approach

- Data Collection
  - Importing the data and inspecting the data frame created initially.
  - Gathering the statistics related to dataframe namely shape, description of rows and columns etc.
- Data Cleaning
  - Checking for null value columns in the dataset.
  - Dropping columns with large count of NULL values.
  - Dropping unnecessary columns.
  - Imputation of NaN values in columns like ‘Select’, ‘Last Activity’ etc.

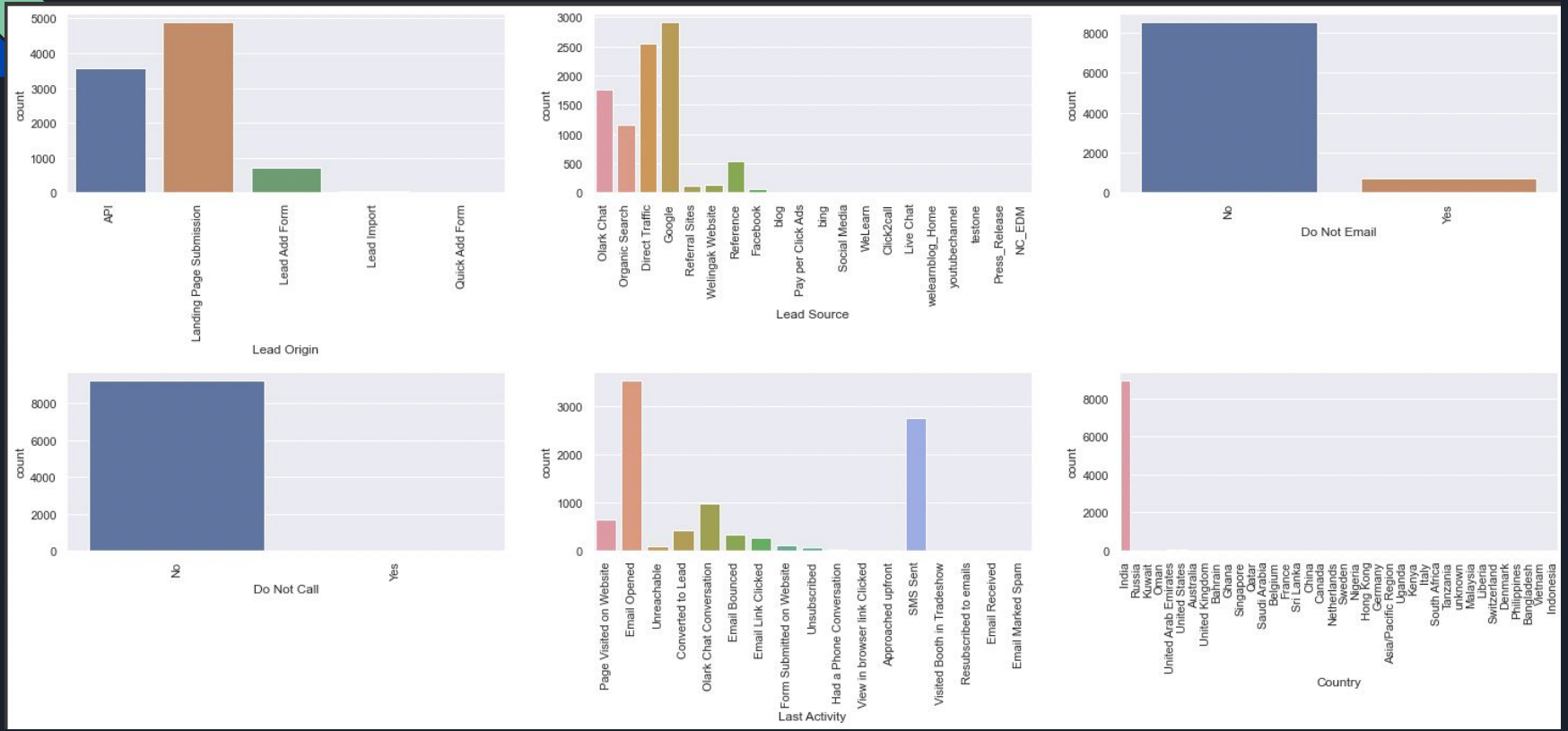
- Exploratory Data Analysis (EDA)
  - Checking for outliers.



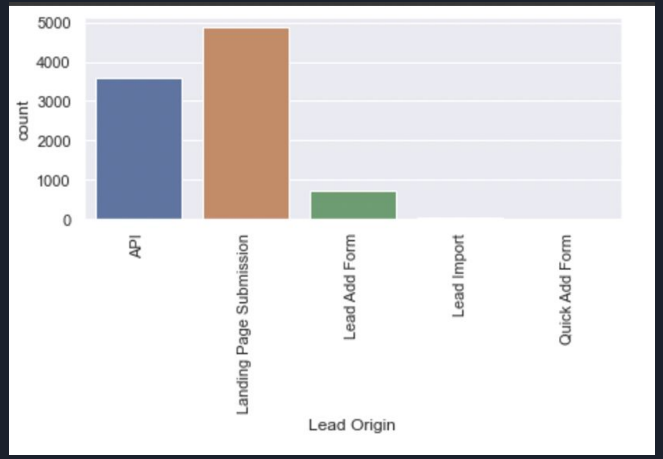
- Correlation analysis.

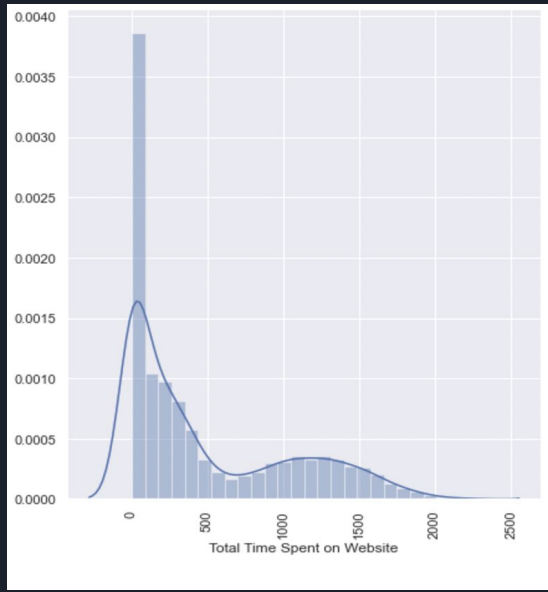


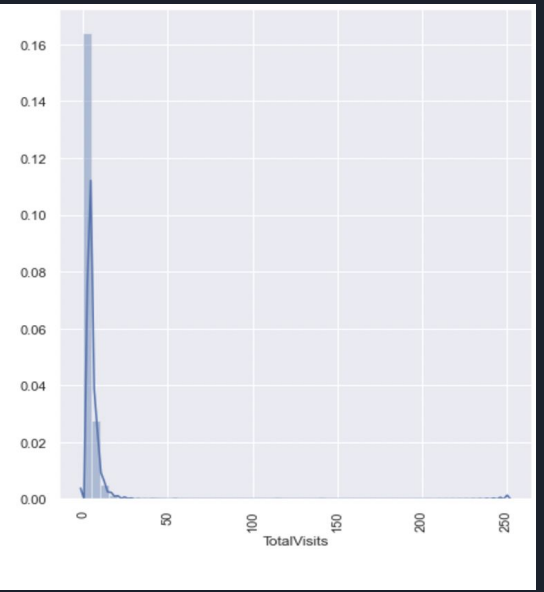
- Countplot to check data distribution.



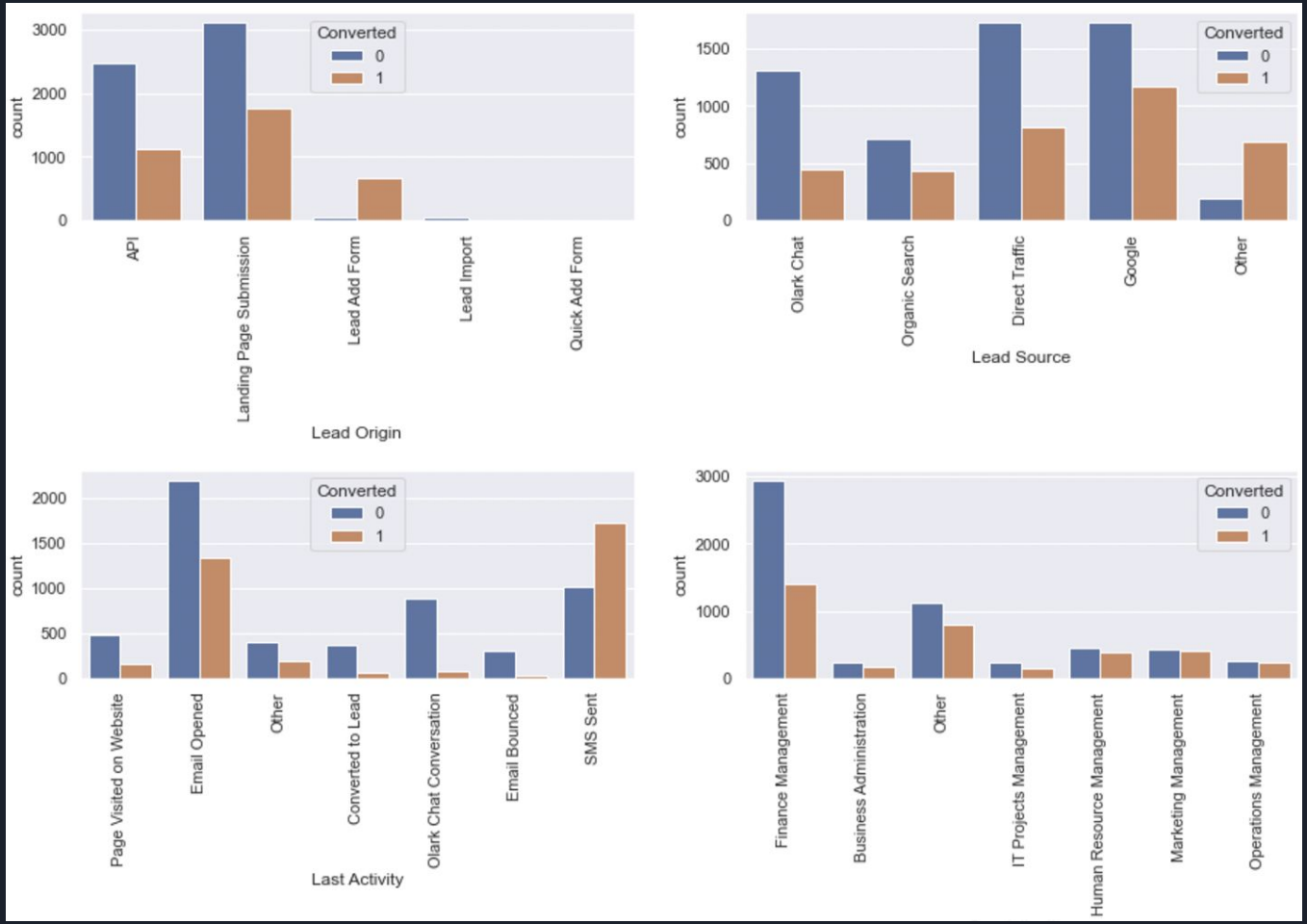
- Univariate Analysis - Categorical and Continuous.





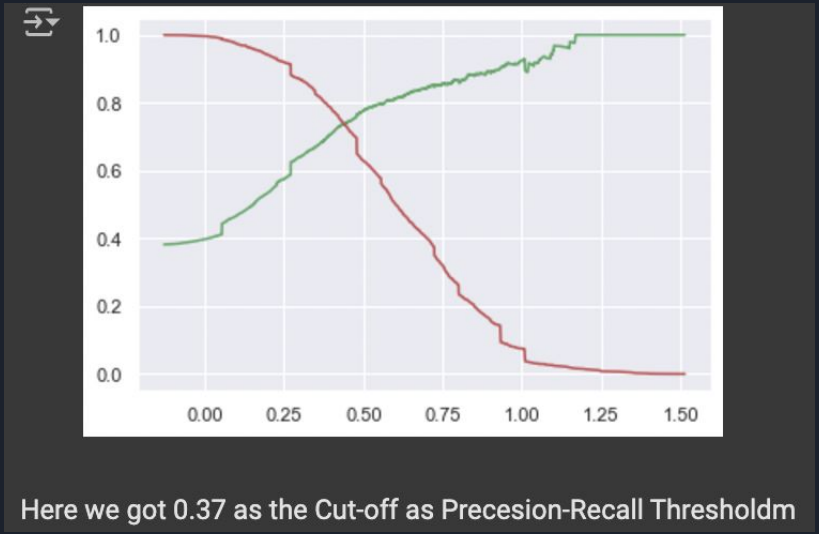


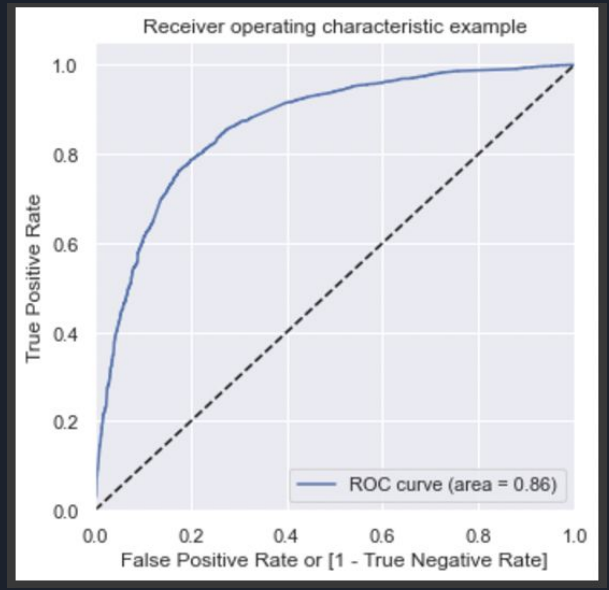
- Bivariate Analysis



- Dummy Variable Creation
  - Creating dummies for categorical variables.
- Data Preparation and Feature Scaling
  - Splitting dataset into test and training datasets for model evaluation.
  - Scaling and transforming datasets for standardisation.
- Model Building and training
  - Creating a Logistic Regression model.
  - Training the model using training dataset.
  - Model building using different VIF values and p-values by dropping non required columns.

- Model Evaluation
  - Making predictions on training dataset.
  - Checking for accuracy, confusion matrix , precision and other parameters.
  - Graphs plotted between various output parameters.
- Test Dataset Prediction
  - Making predictions on test dataset.
  - Checking for accuracy, confusion matrix , precision and other parameters.





Results

We can conclude the following that the variables that important the most in the potential buyers are:

- The total time spent on the Website.
- Total number of visits.
- When the lead source was: a. Google b. Direct traffic c. Organic search d. Olark Chat
- When the last activity was: a. SMS b. Olark chat conversation When the lead origin is Lead add format.

Training Dataset

Accuracy: 78.57%

Sensitivity: 81.02%

Specificity: 77.06%

Test Dataset

Accuracy: 69.84%

Sensitivity: 27.85%

Specificity: 97.25%