Lead Scoring Case Study

By : Pramai Das

Pramai Das

Data Analyst

Agenda

01 **Introduction**

02 **Exploratory Data Analysis**

03 **Model Building**

04 **Summary**

X EDUCATION

1010011010010000010101
0011110111011011011010
1010000111001010 11001
0101001110101000010101
0001011010110110 11010
0010101110001010 10001
0100010111010110 00100
1101001101001000 01010
1001111011101101101101
0101000011001010 1100
1010100111010100 01010
1000101101011011 01101
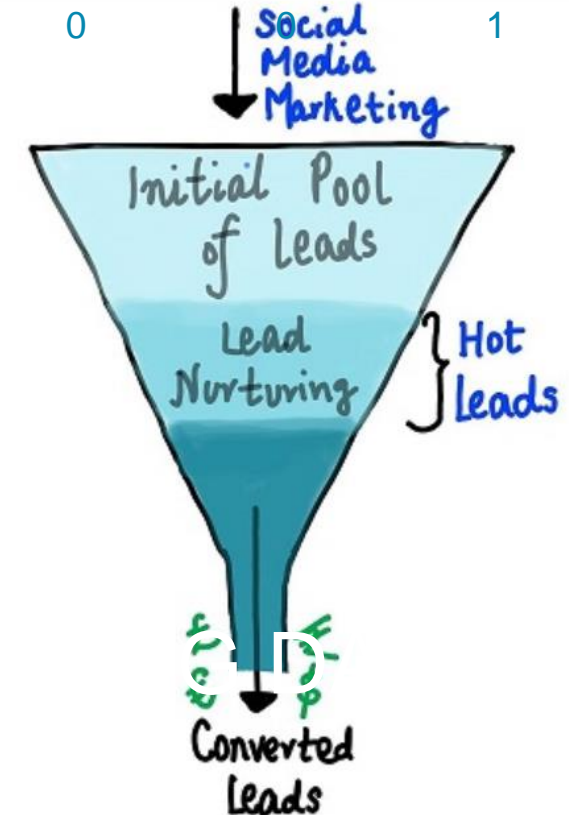0   1

An Education Company

X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

Objective

The company requires us to build a model wherein you need to assign a lead score 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 70%.

Social Media Marketing

Initial Pool of Leads

Lead Nurturing } Hot Leads

Converted Leads

Lead Conversion Process - Demonstrated as a funnel

Basic Architecture

X Education Data

A leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.

Predictive Analysis

Lead Score

A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Logistic Regression Model

Building 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.

Data Understanding

Data understanding is the knowledge you have about data, the needs the data will satisfy, its content and location.

Basic Checks

You can simply impress your audience and add a unique zing and appeal to your Presentations. Easy to change colors, photos and Text.

EDA

Exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.
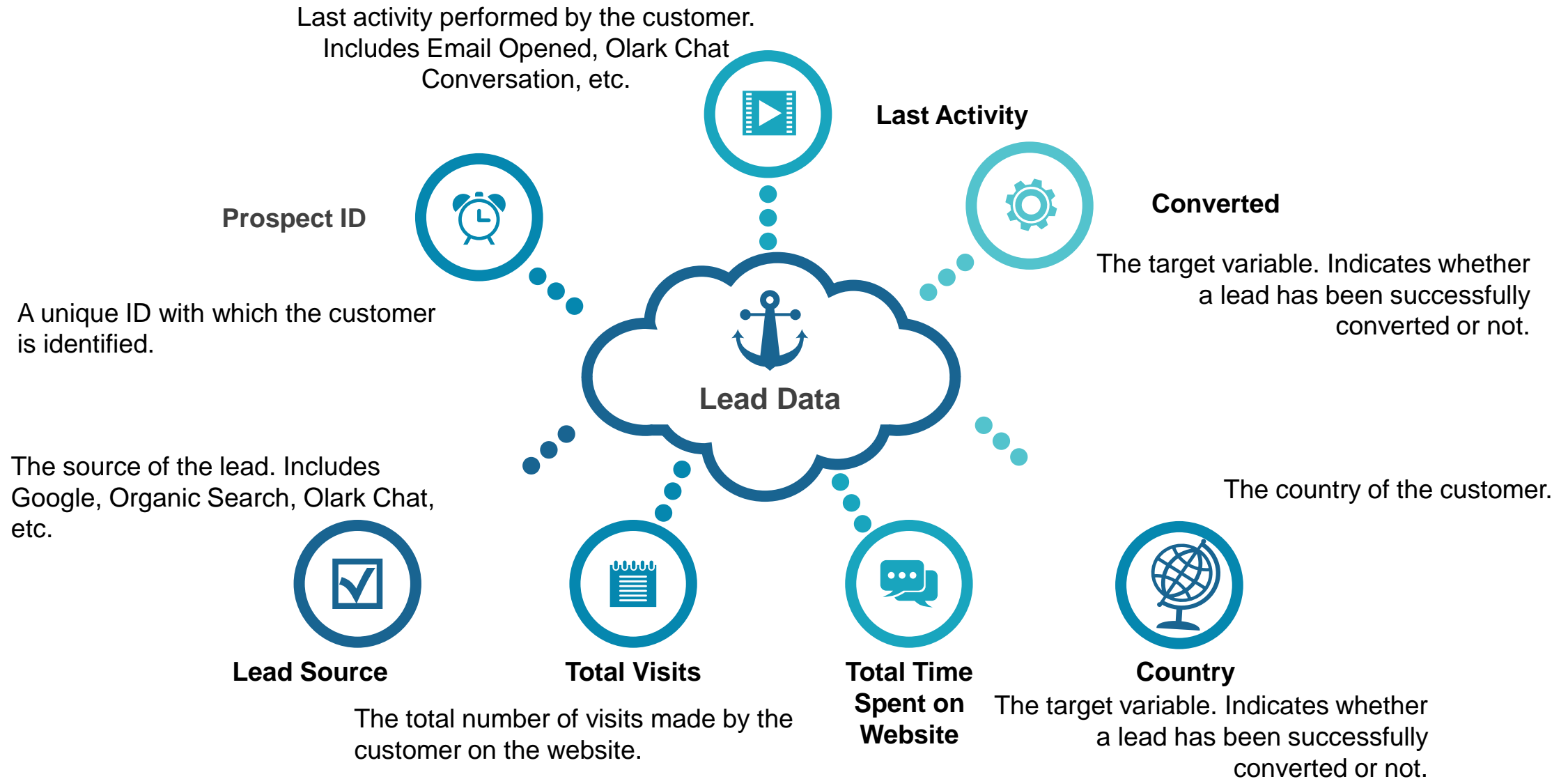
Model building

Using algorithms that iteratively learn from data, machine learning models facilitate computers to find hidden insights from Big Data without being explicitly programmed where to look.

Predictive Analysis

Encompasses a variety of statistical techniques from predictive modelling, and machine learning, that analyze current and historical facts to make predictions about future or otherwise unknown events.
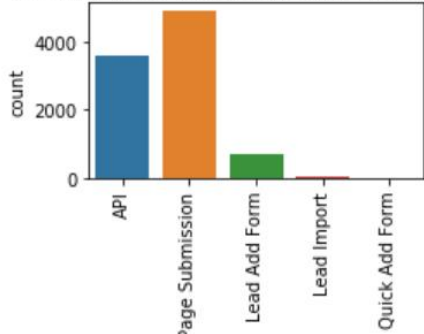
Basic Flow

Analyzing the Dataset

Last activity performed by the customer.
Includes Email Opened, Olark Chat
Conversation, etc.

Last Activity

Prospect ID

Converted

The target variable. Indicates whether
a lead has been successfully
converted or not.

A unique ID with which the customer
is identified.

Lead Data

The source of the lead. Includes
Google, Organic Search, Olark Chat,
etc.

The country of the customer.

Lead Source

Total Visits

**Total Time
Spent on
Website**

Country

The total number of visits made by the
customer on the website.

The target variable. Indicates whether
a lead has been successfully
converted or not.

Visualizing the Data

Plotting Lead Origin

Plotting data for target in terms of total count

The graph implies that the customer do open their email a lot as the count being highest among all.

Plotting Last Activity

Plotting data for target in terms of total count

Visualizing the categorical variable

From the lead origin graph we can clear say there is a good variation and **Landing page submission** has the highest count

Plotting Lead Source

Plotting data for target in terms of total count
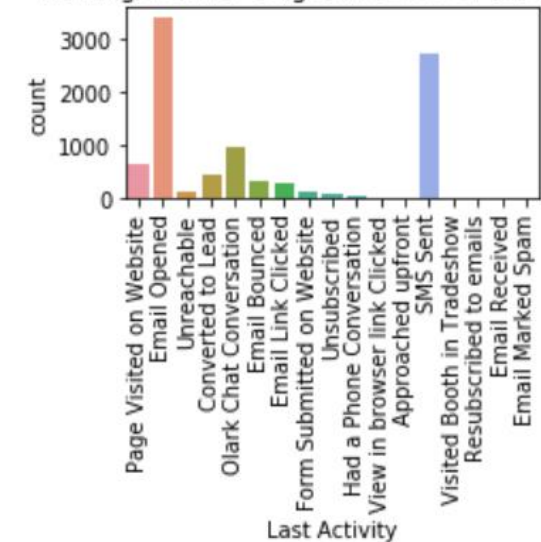
Plotting What is your current occupation

Plotting data for target in terms of total count

From the lead score graph we can clear say there is a good variation and most customer are using the **Google** as their lead source

Unemployed Customer have the highest count in occupation category
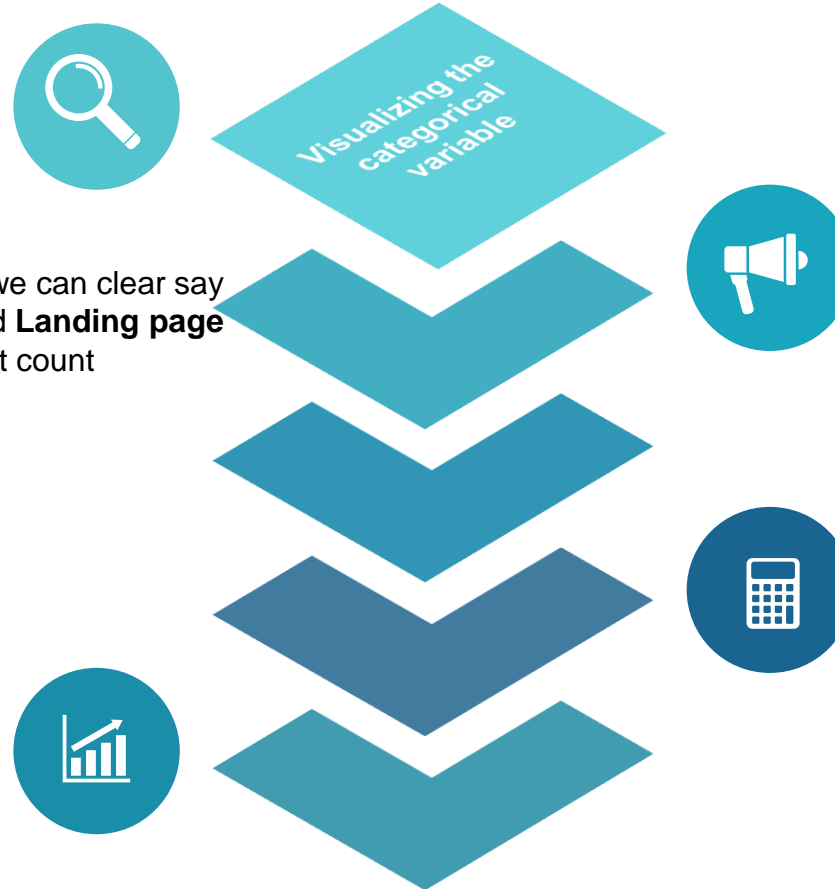
Visualizing the Data
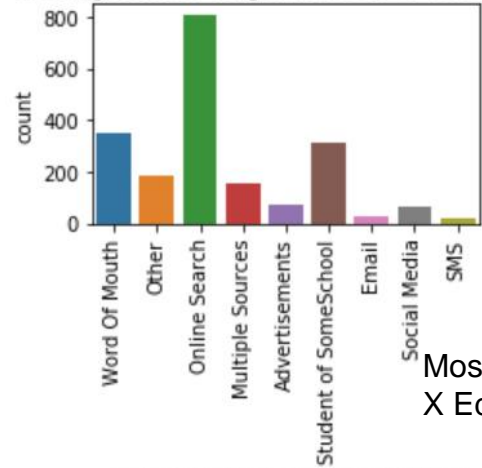
Most of the customer have heard about the X Education form online search
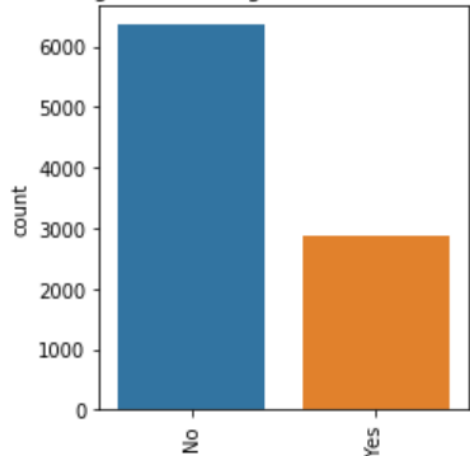
The last activity by customer is that the maximum customer has modified their data.

From the lead score graph we can clear say there is a good variation and most customer are using the **Google** as their lead source
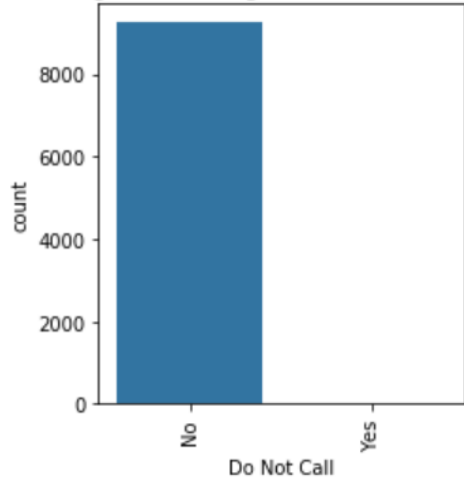
Maximum Customer have chosen not to email about the course

Visualizing the Data

Plotting Do Not Call

Plotting data for target in terms of total count

All the below graphs does not have any variation. Thus considering this attribute is not an ideal choice.

Plotting Search

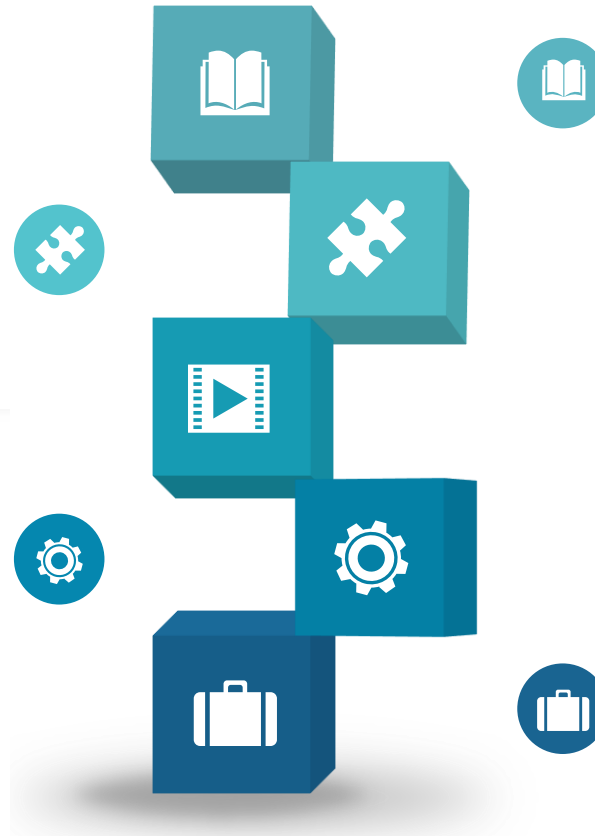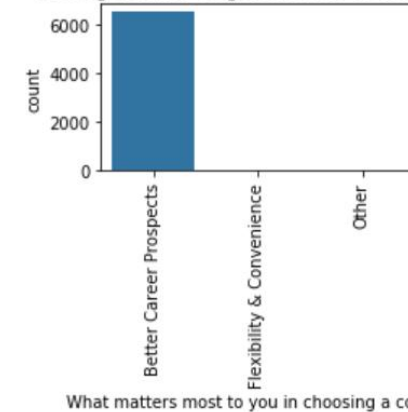Plotting data for target in terms of total count
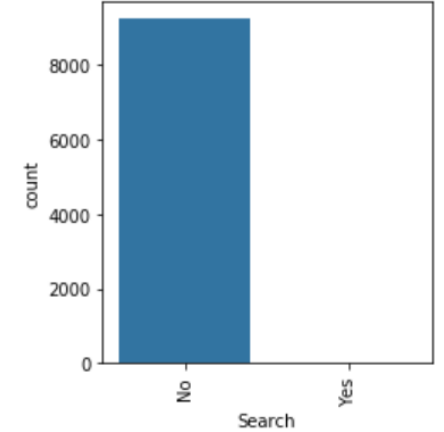
Plotting Update me on Supply Chain Content

Plotting data for target in terms of total count
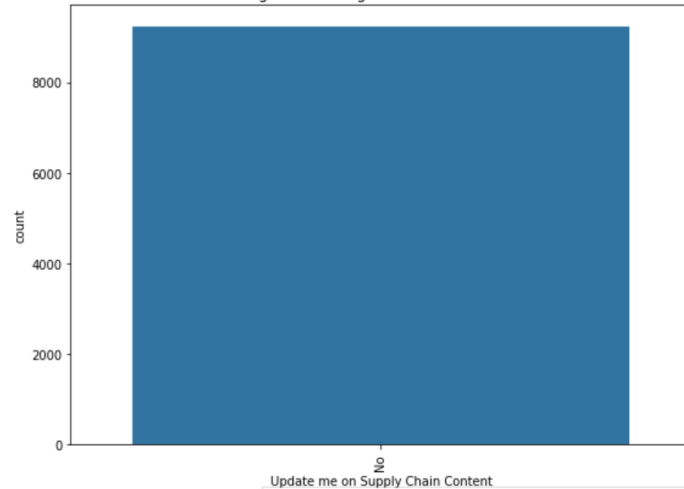
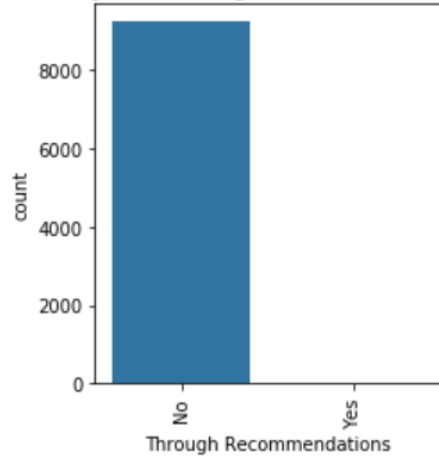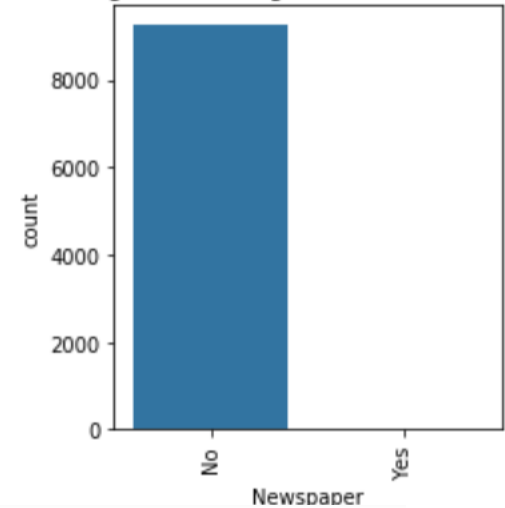Plotting What matters most to you in choosing a course

Plotting data for target in terms of total count

Visualizing the Data

Plotting data for target in terms of total count

All the below graphs does not have any variation. Thus considering this attribute is not an ideal choice.
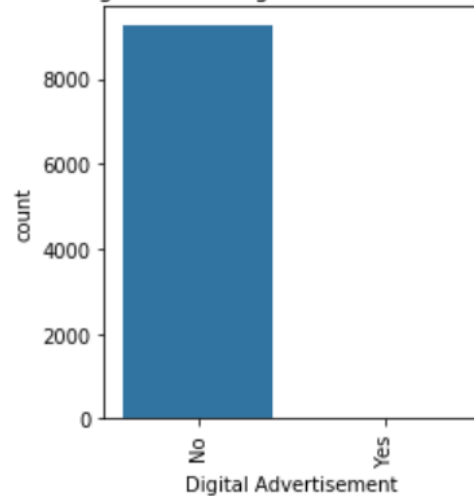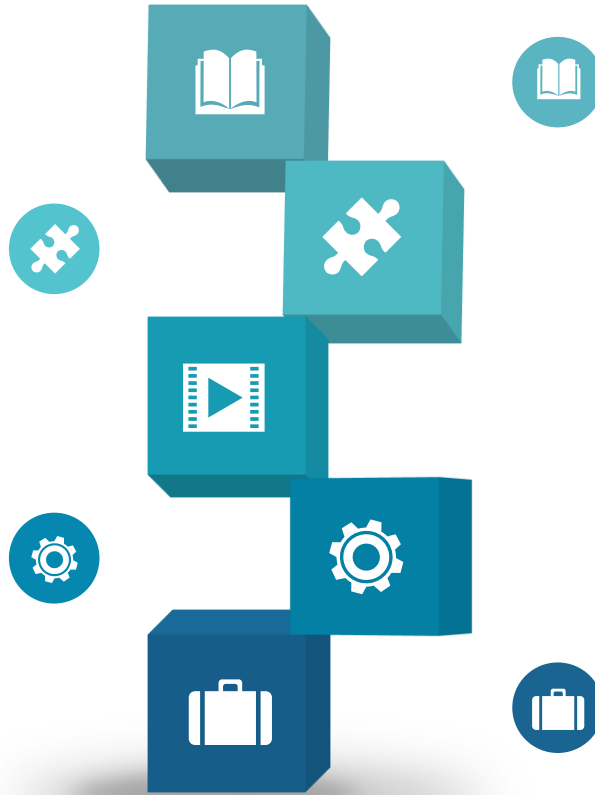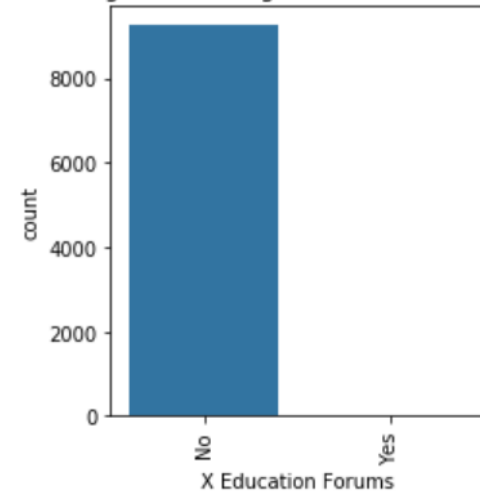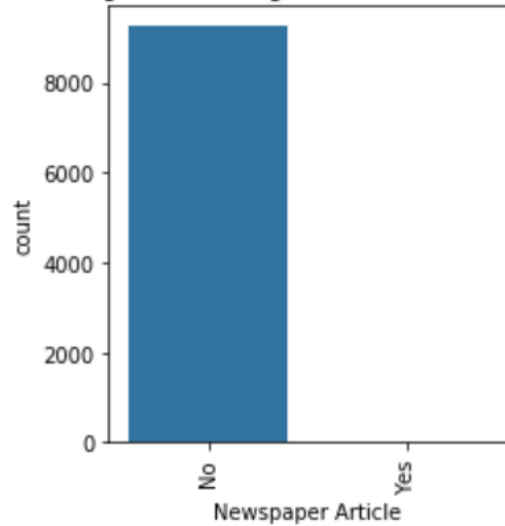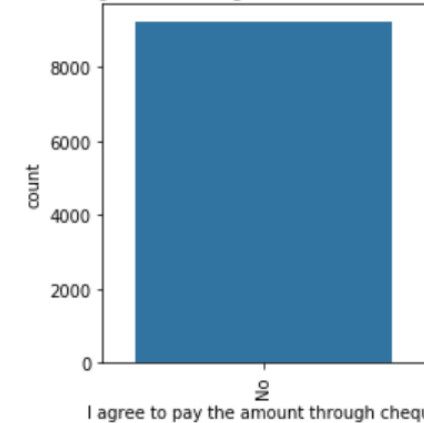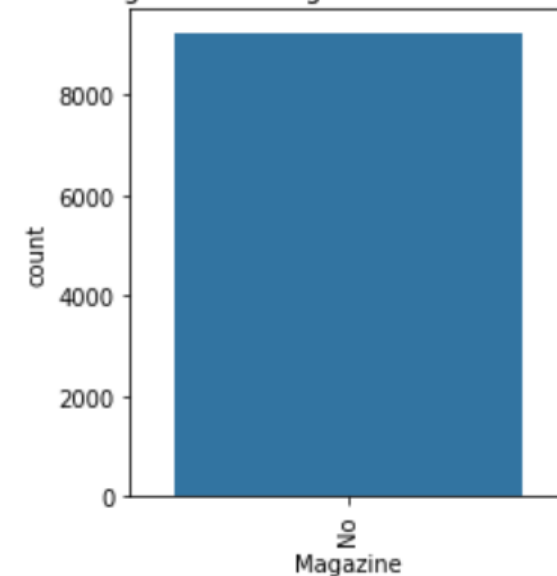
Plotting data for target in terms of total count

Plotting data for target in terms of total count

Plotting data for target in terms of total count

Visualizing the Data

All the below graphs does not have any variation. Thus considering this attribute is not an ideal choice.

Plotting Newspaper Article

Plotting data for target in terms of total count

Plotting I agree to pay the amount through cheque
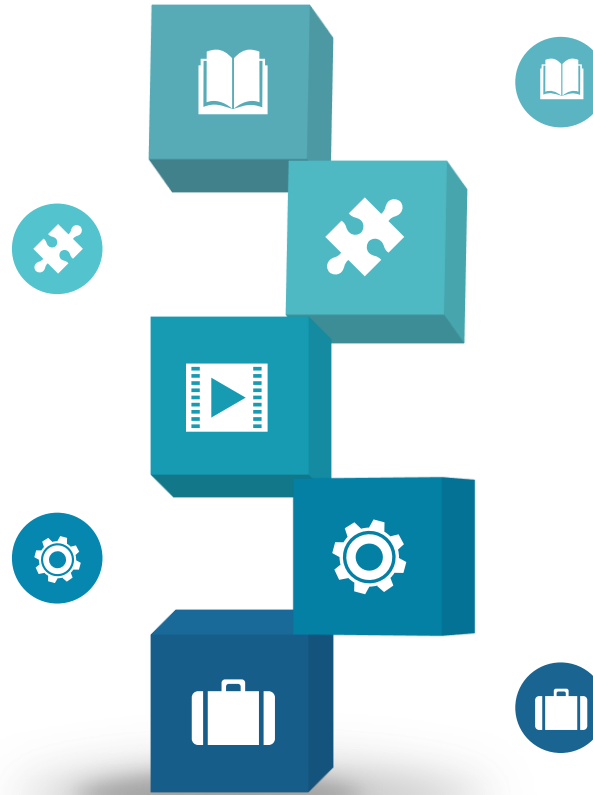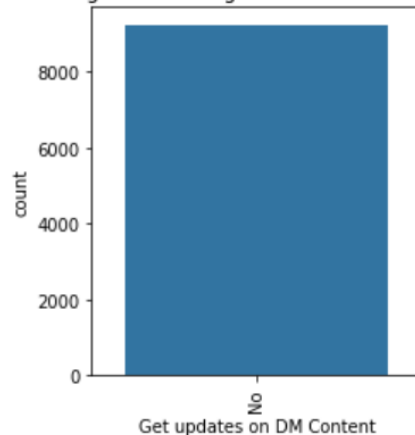
Plotting data for target in terms of total count

Plotting Get updates on DM Content

Plotting data for target in terms of total count

Plotting Magazine

Plotting data for target in terms of total count

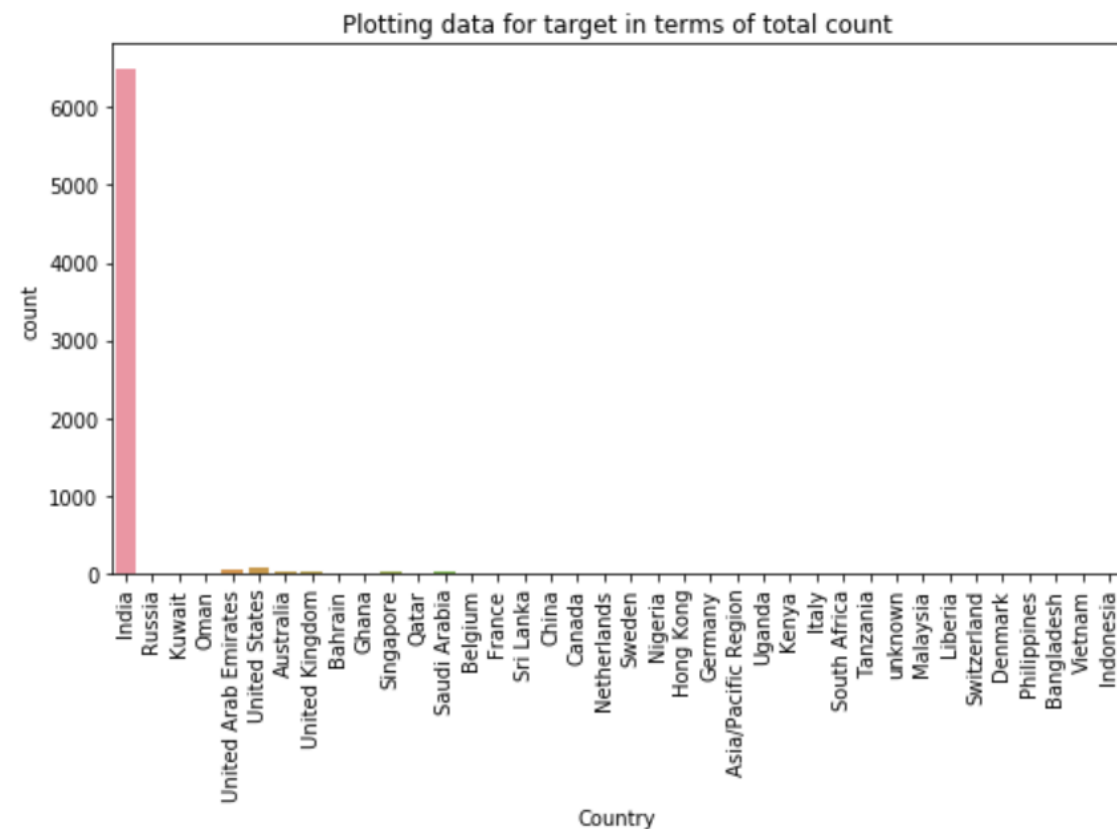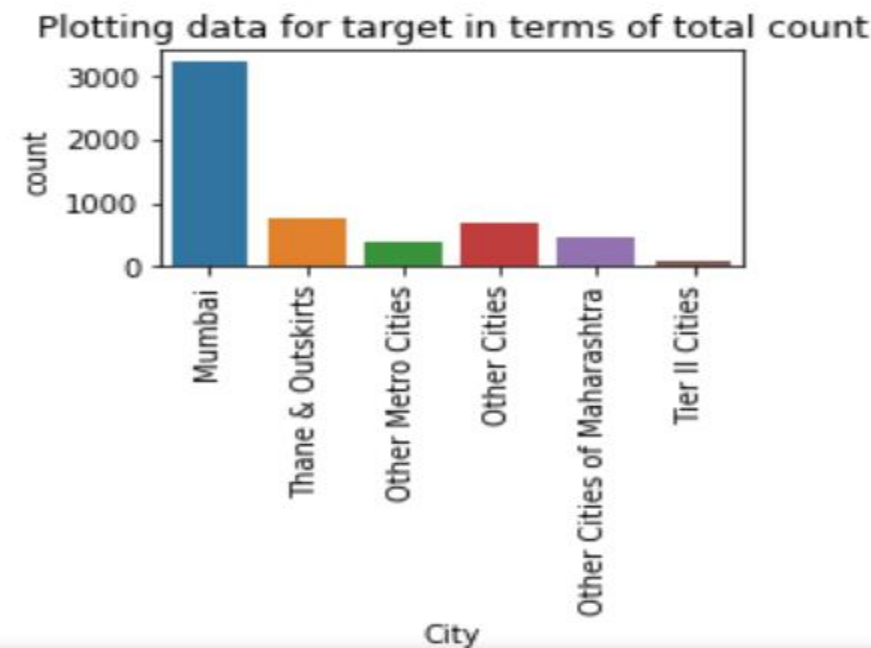# Visualizing
the County and City

County : Maximum customers are from India

City : Maximum customers are from Mumbai

Class Imbalance

Imbalanced classes put "accuracy" out of business. This is the most common problem in machine learning (specifically in classification), occurring in datasets with a disproportionate ratio of observations in each class.
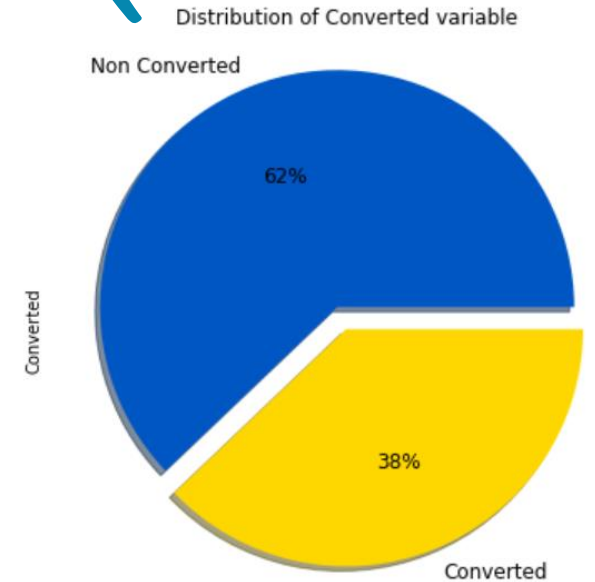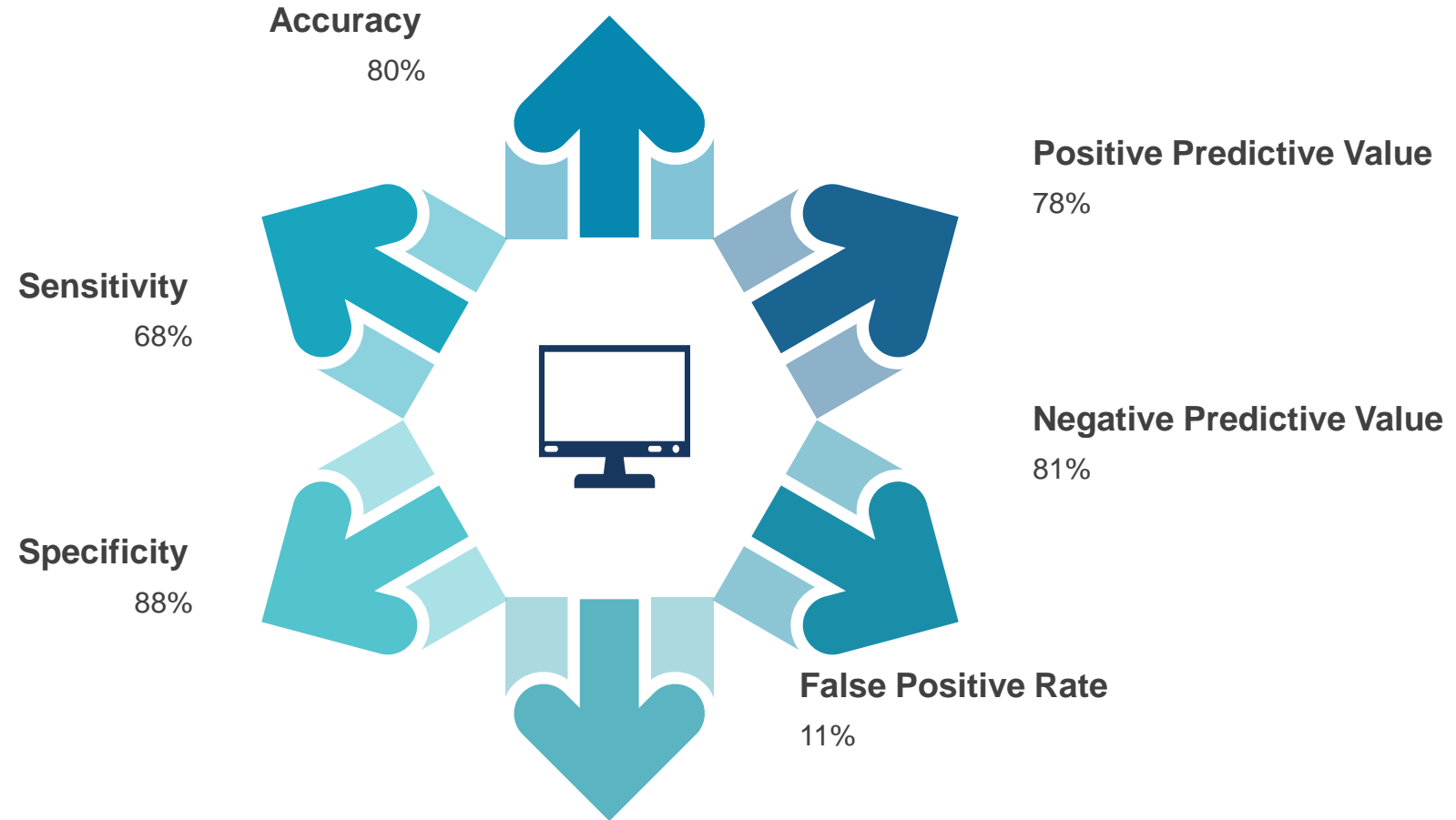
Non Converted rate comes out to be 62%
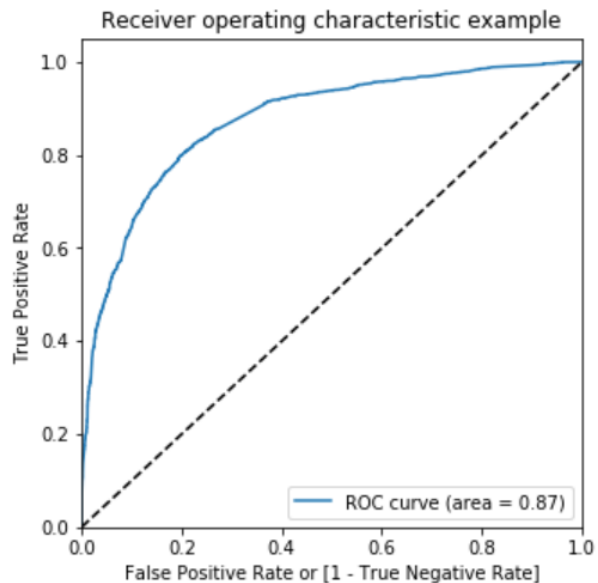
Converted rate comes out to be 38%
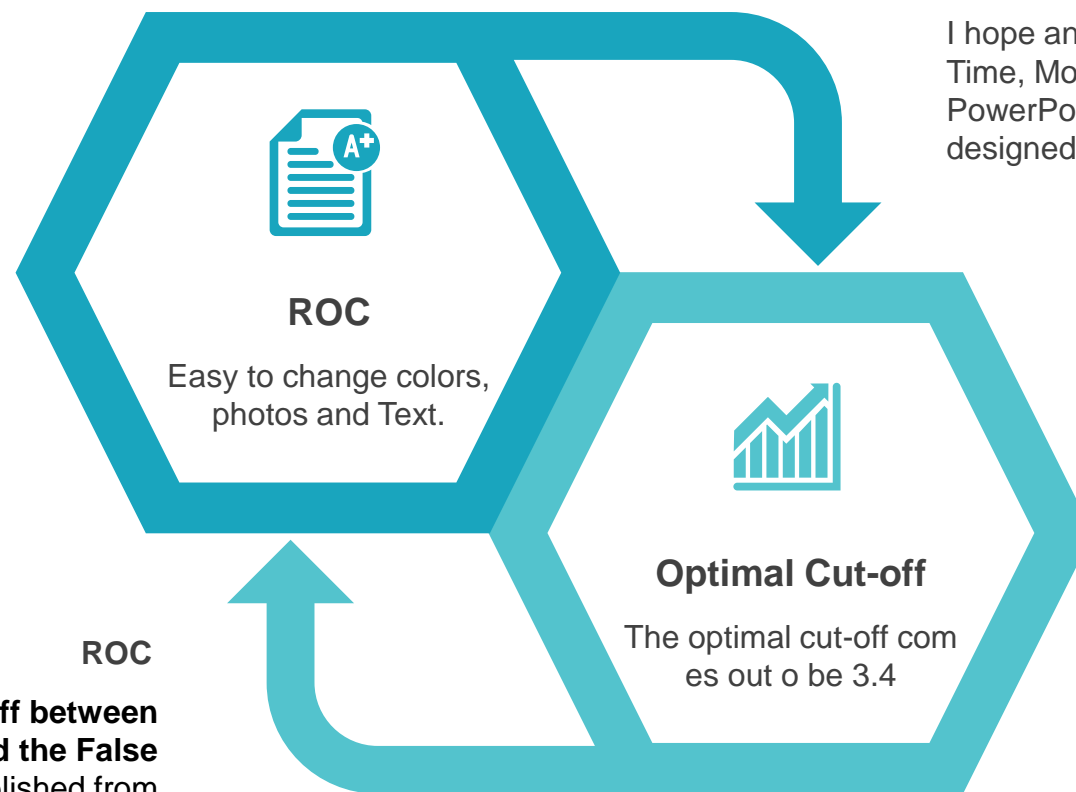
Thus we can conclude that we have good class imbalance.

Distribution of Converted variable

Non Converted

Converted

62%

38%

Converted

Model Result

ROC And Optimal Cut-off

Receiver operating characteristic example

ROC

Easy to change colors, photos and Text.

Optimal Cut-off

The optimal cut-off comes out o be 3.4

Optimal Cut-off

I hope and I believe that this Template will your Time, Money and Reputation. Get a modern PowerPoint Presentation that is beautifully designed.
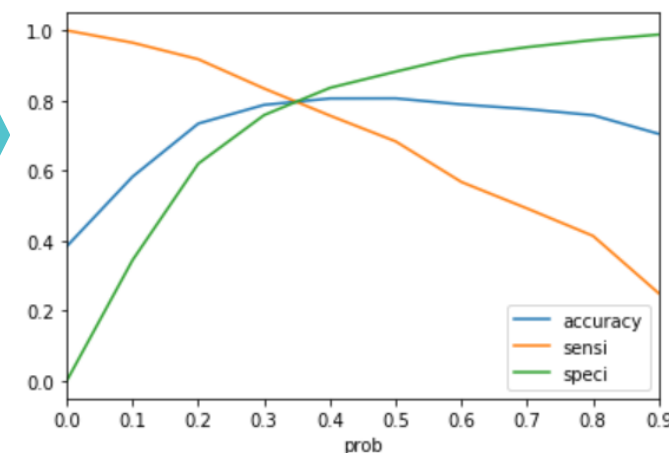
ROC

ROC Curves which show the **tradeoff between the True Positive Rate (TPR) and the False Positive Rate (FPR)**. And as was established from the formulas above, TPR and FPR are nothing but sensitivity and (1 - specificity), so it can also be looked at as a tradeoff between sensitivity and specificity.
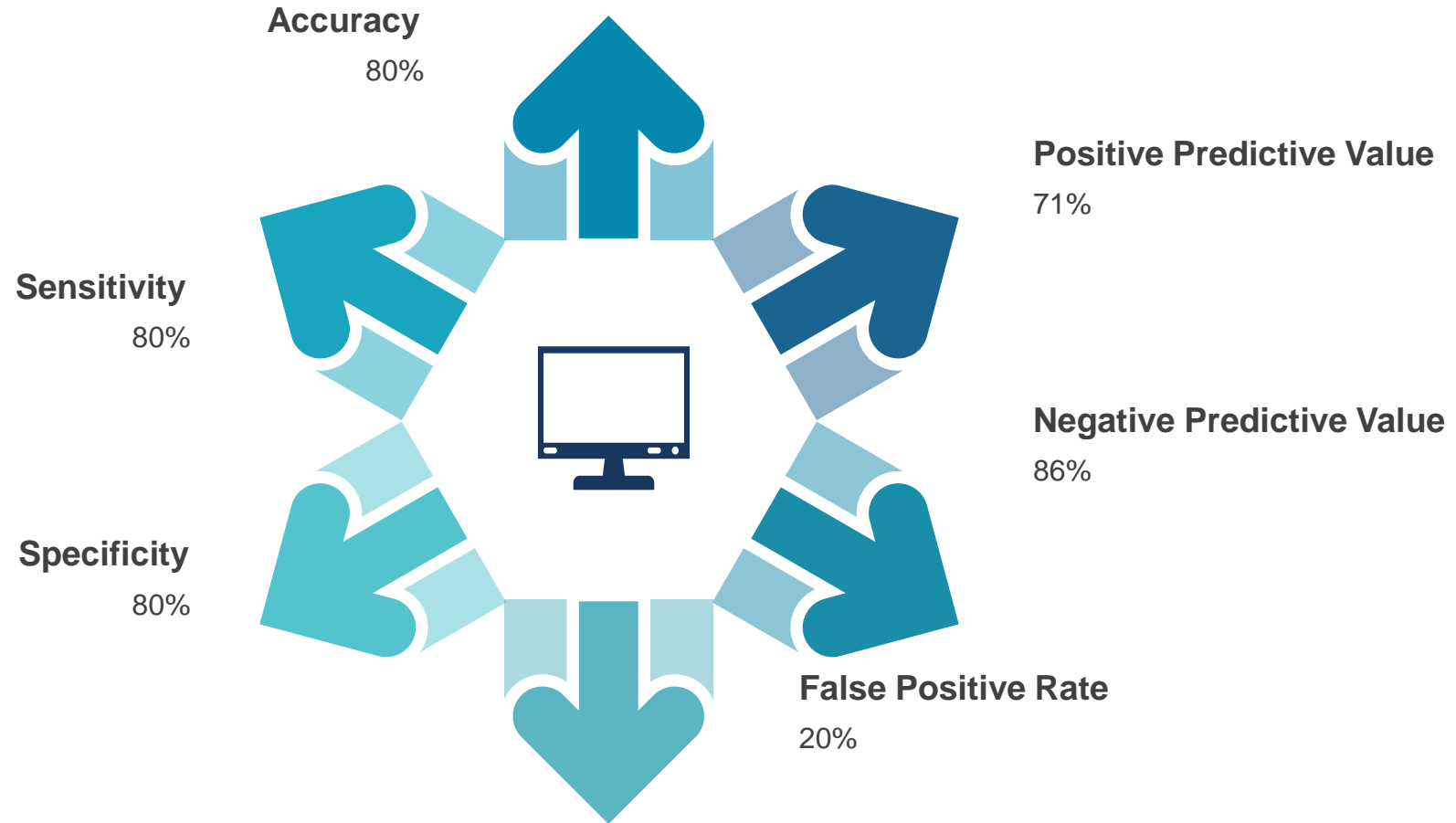
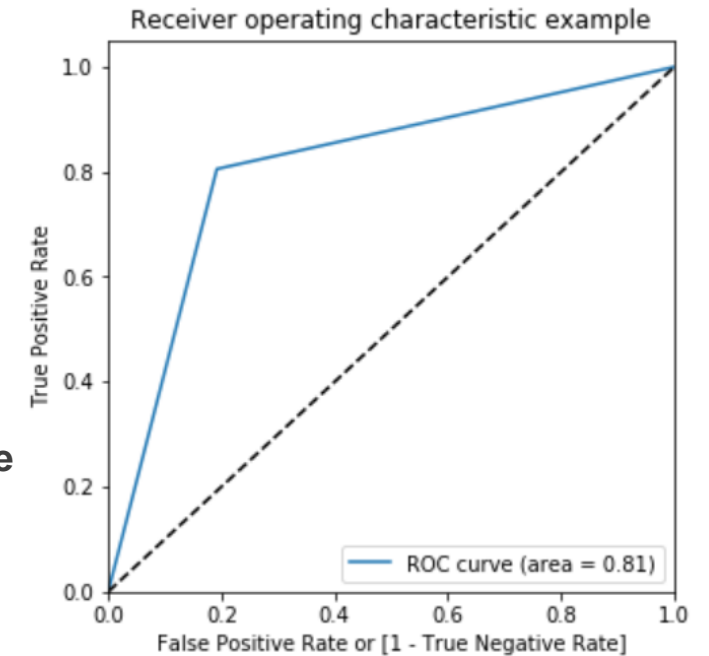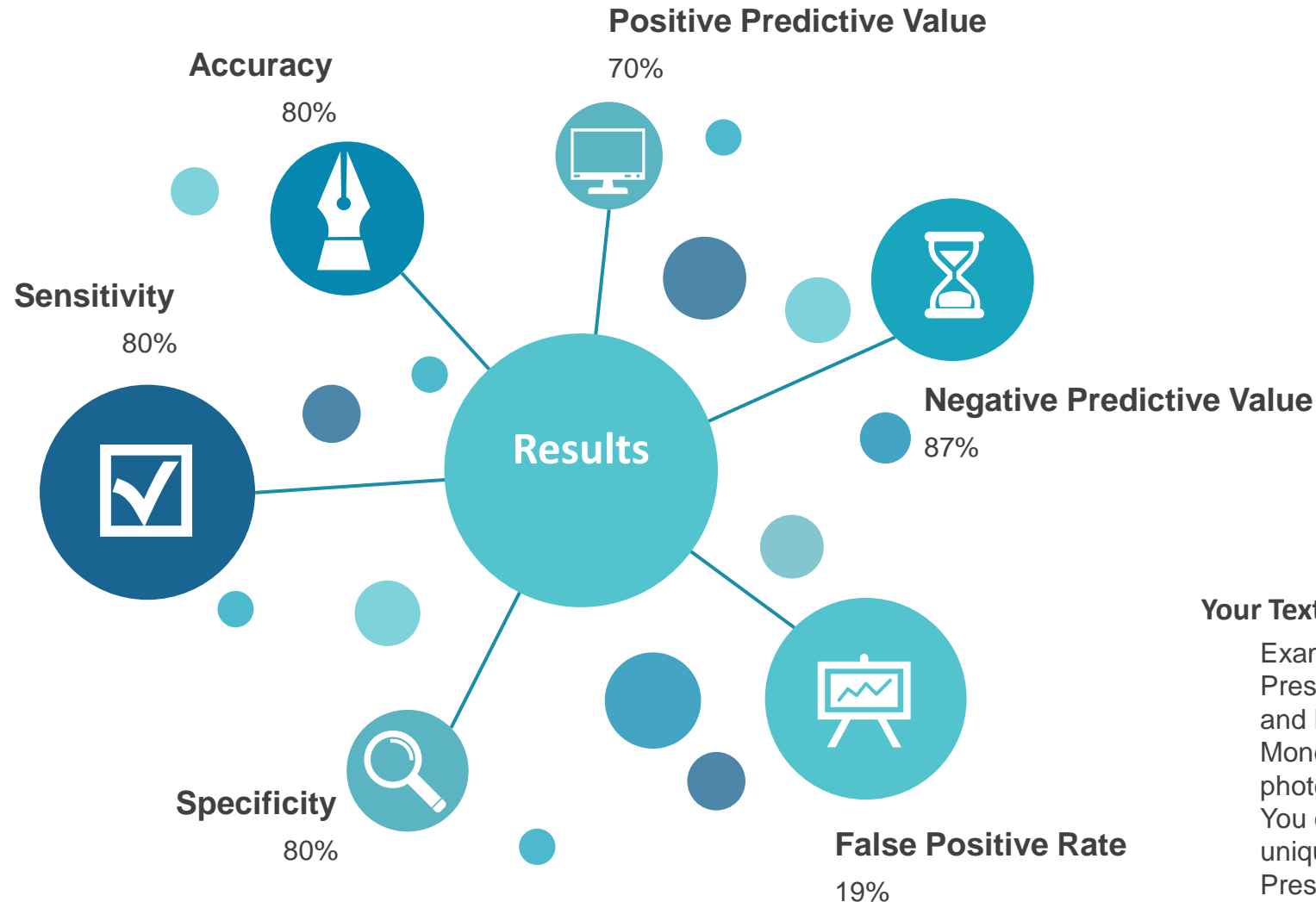# Model Result based on optimal threshold

Accuracy
80%

Positive Predictive Value
71%

Sensitivity
80%

Negative Predictive Value
86%

Specificity
80%

False Positive Rate
20%

Predictive Results

Accuracy
80%

Positive Predictive Value
70%

Sensitivity
80%

Results

Negative Predictive Value
87%

Specificity
80%

False Positive Rate
19%

Receiver operating characteristic example

True Positive Rate

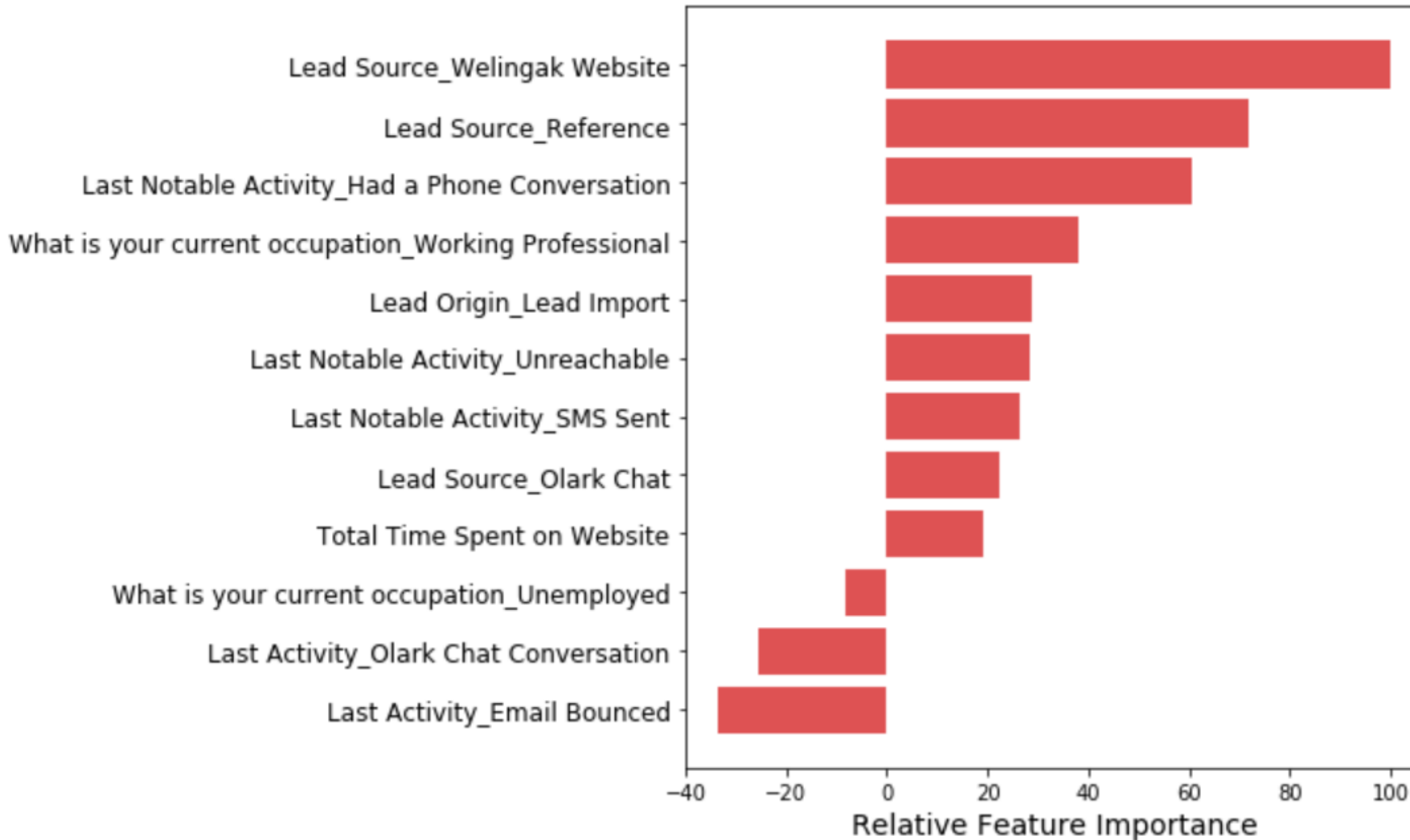False Positive Rate or [1 - True Negative Rate]

ROC curve (area = 0.81)

Your Text Here

Example Text : Get a modern PowerPoint Presentation that is beautifully designed. I hope and I believe that this Template will your Time, Money and Reputation. Easy to change colors, photos and Text.
You can simply impress your audience and add a unique zing and appeal to your Reports and Presentations with our Templates. Get a modern PowerPoint Presentation that is beautifully designed.

Top Features

THANK YOU