Lead Scoring Case Study

GOAL : THE COMPANY WANTS TO KNOW THE MOST PROMISING LEADS I.E. THE LEADS THAT ARE MOST LIKELY TO CONVERT INTO PAYING CUSTOMERS.

Submitted By:
1. Vaishali Papneja
2. Sushasree Vasudevan Suseel Kumar

Business Understanding

An Education Company sells Online courses to Industry professionals. When these customers land on the website and enter the information like emails, phone numbers etc. they are classified as a “Lead”. The current Lead Conversion Rate is poor.

Conversion Rate : Total number of conversions divided by the number of leads.

We need to identify the potential leads (leads that are most likely to convert into paying customers) by building a model wherein we need to assign a lead score to each lead.

The Target Lead Conversion rate should be around 80%

Steps Involved in Analysis

We are a given a dataset from the past with around 9000 data points.

- **Data Cleaning and Data Preparation. Use EDA to analyze the patterns present in the data to check if the features are required for further analysis. Divide the cleaned ,prepared and scaled dataset into train and test dataset.**

We are using Logistic Regression to build the model

- **Build the model using the TRAIN dataset.**
- **Test the model using the TEST dataset.**
- **Evaluate the Model.**
- **Find the Lead Score, Conversion Ratio.**

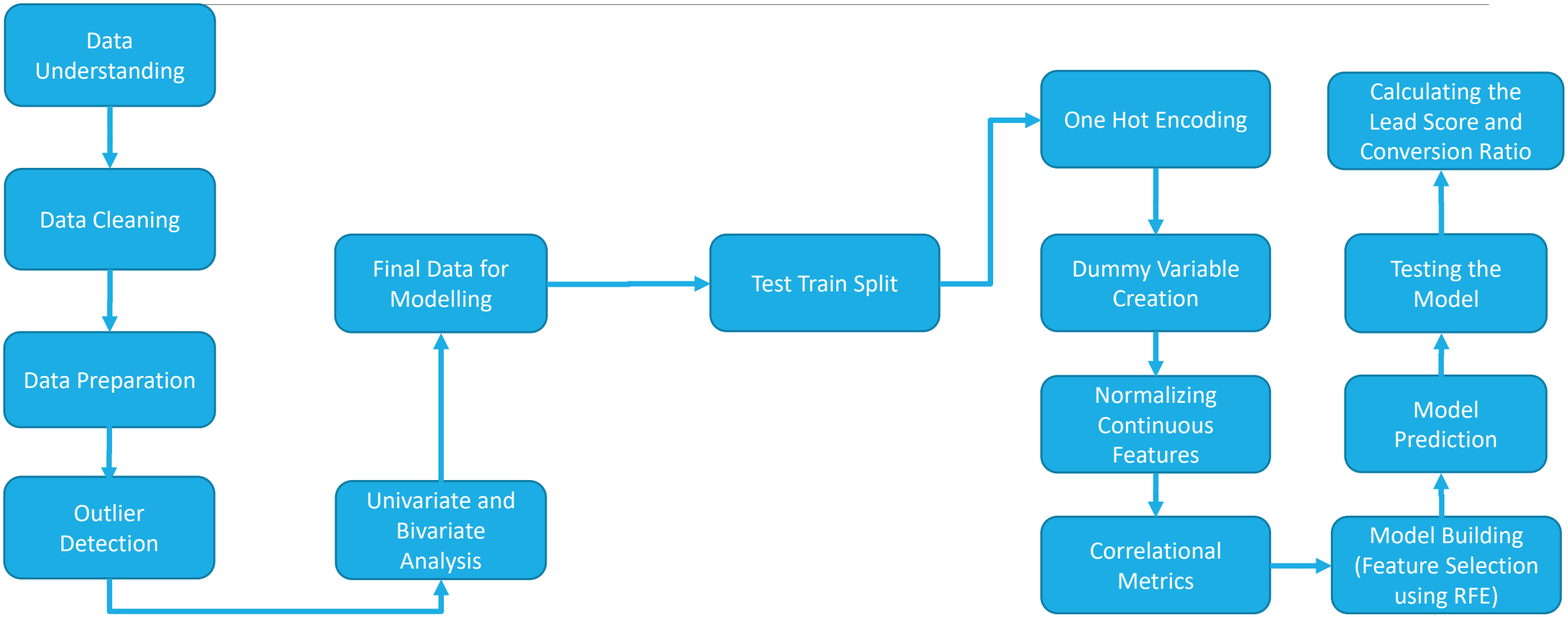
- Approach used,
  - **To find the probability of conversion**
    - In the dataset, Find the probability of conversion on the entire data for visualizations during EDA
      - - Probability of lead Conversion = Converted/(converted + not converted)
      - - Probability of no Lead conversion = not converted/(converted + not converted)
  - **To find the Lead Score,**
    - Lead Score = 100 * Conversion Probability
  - **To Find the Conversion Ratio**
    - With the help of conversion probability and cutoff value
      - Ratio : Sum of converted in the dataset/length of the dataset

Assumptions

In the dataset, we have many columns which has the value as Select. This is because the lead/employee did not select an option while filling the form.

So replacing all the Select values as NULL.

Problem Solving Methodology



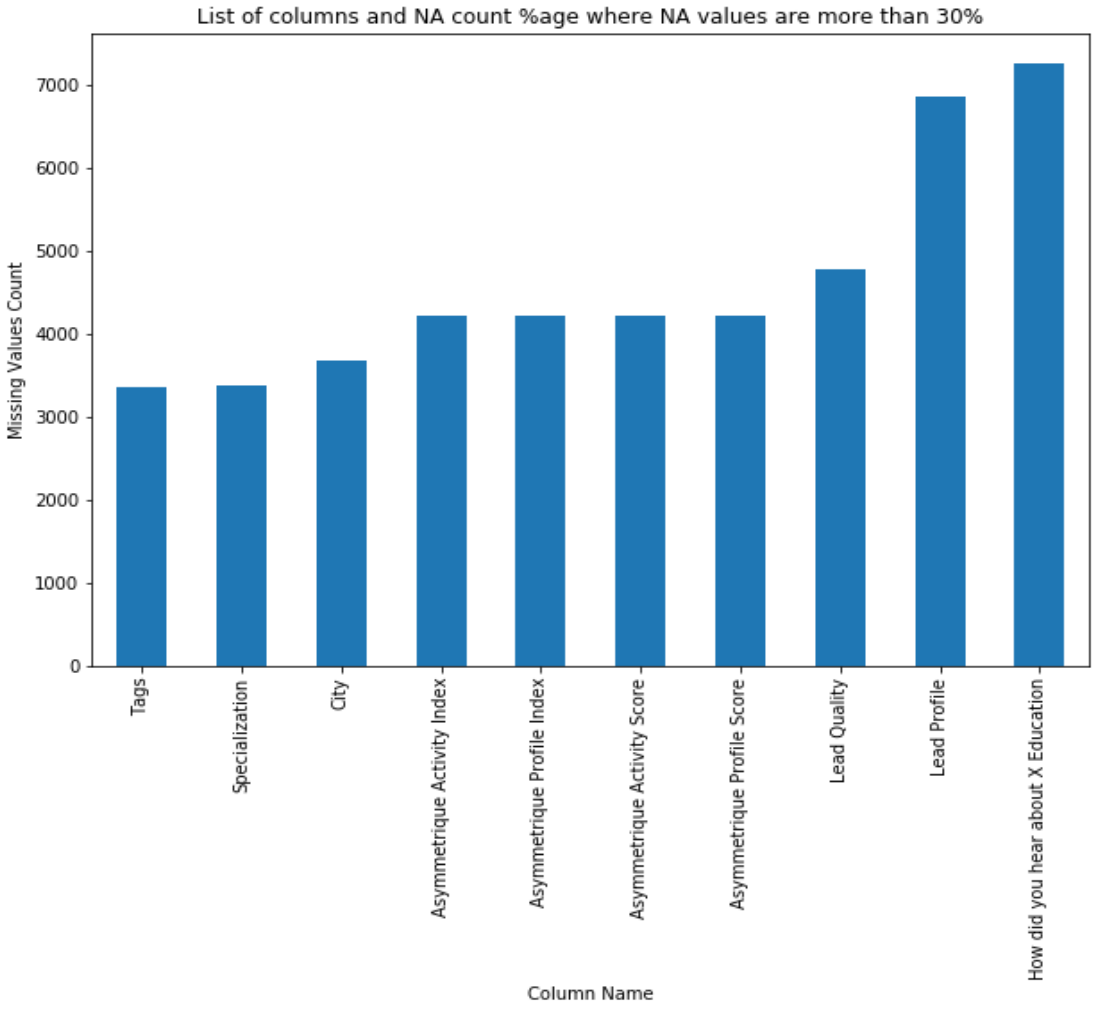
Data Cleaning and Data Preparation

Manage Missing Data in the Lead Dataset

- **We have 10 columns which have more than 30 % missing values.**

There are various ways to deal with missing values. Out of which, most common methods are as below:

- Remove those columns if we have higher proportion of missing data
- Replace them with
  - Mean/Median/Mode in case of quantitative variables.
    - Replace them with mean if data in that filed is distributed normally.
    - Replace them with median if there are outliers present in that particular field.
    - Replace with mode if replacing with most repeated value of filed makes sense.
  - Most repeated value in case of categorical variables.
- Replace with a default value
- Leave as it is.
- **In our dataset, we are dropping two columns which have more than 70% missing data as they cannot provide any insight.**
  - **Lead Profile (78%)**
  - **How did you hear about X Education (74%)**



Outliers Detection and Ways to Deal with them

- An outlier is a data point that differ significantly from other data points.
- It may be due to variability in the measurement or may also indicate an experimental error.
- One should be very cautious while dealing with outliers.

For example –

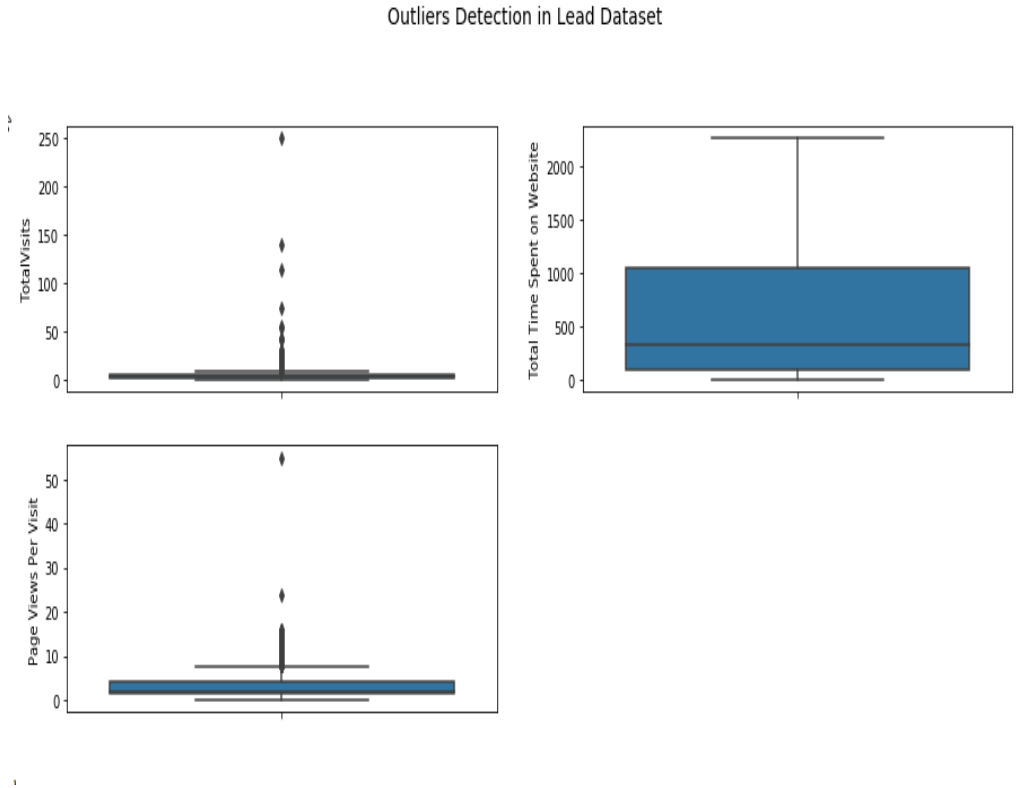
- Outliers introduced due to experimental error can directly be discarded.
- But in the other case, where it has occurred due to variability of measurement, one should take steps to fulfill the purpose.
- Sometimes it is better to remove highly skewed outlier and keep rest of the outliers to get a better insight. While in some other cases, it is better to remove all the outlier data points.

Outliers Detection in Lead Dataset

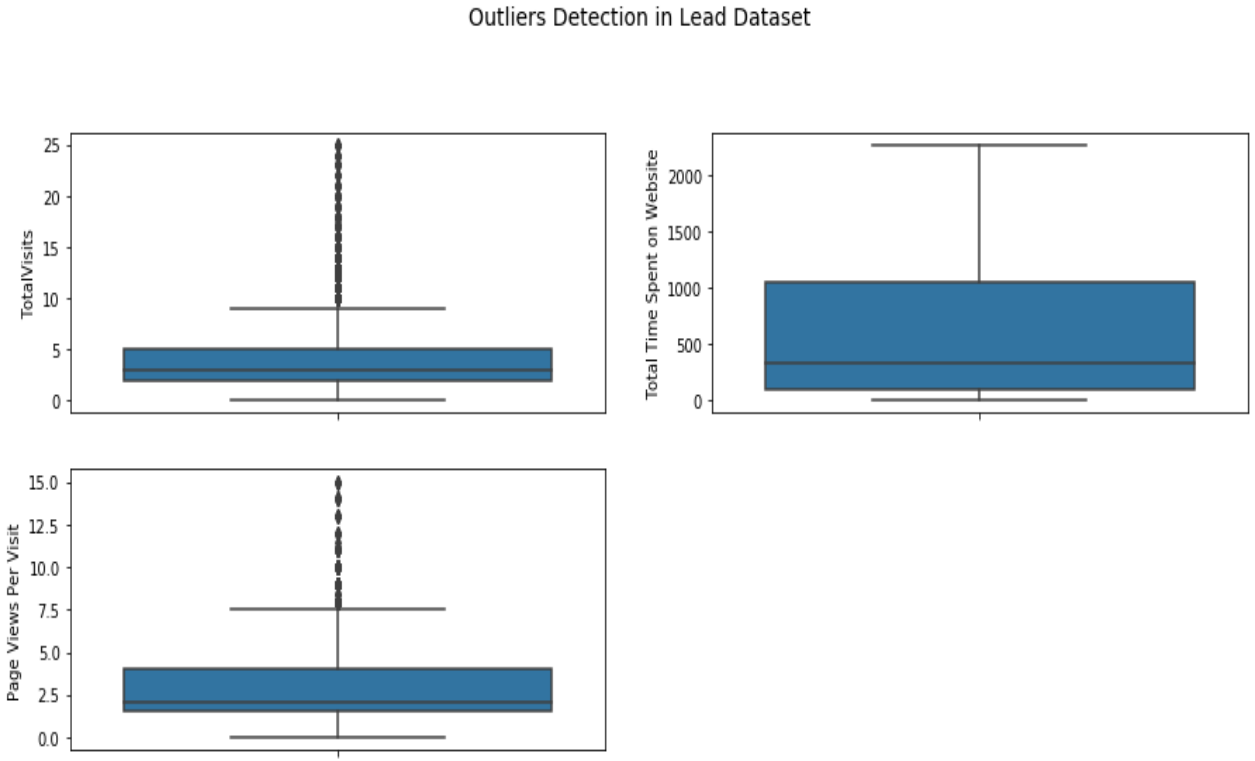
Looking at the box plots, we can see that the dataset has Outliers. Removing Outliers for two columns , Total Visits and Page Views Per Visit.

To remove the Outliers, we will consider the cut off between 5 % and 95 % of the data

DATASET WITH OUTLIERS



DATASET AFTER REMOVING OUTLIERS

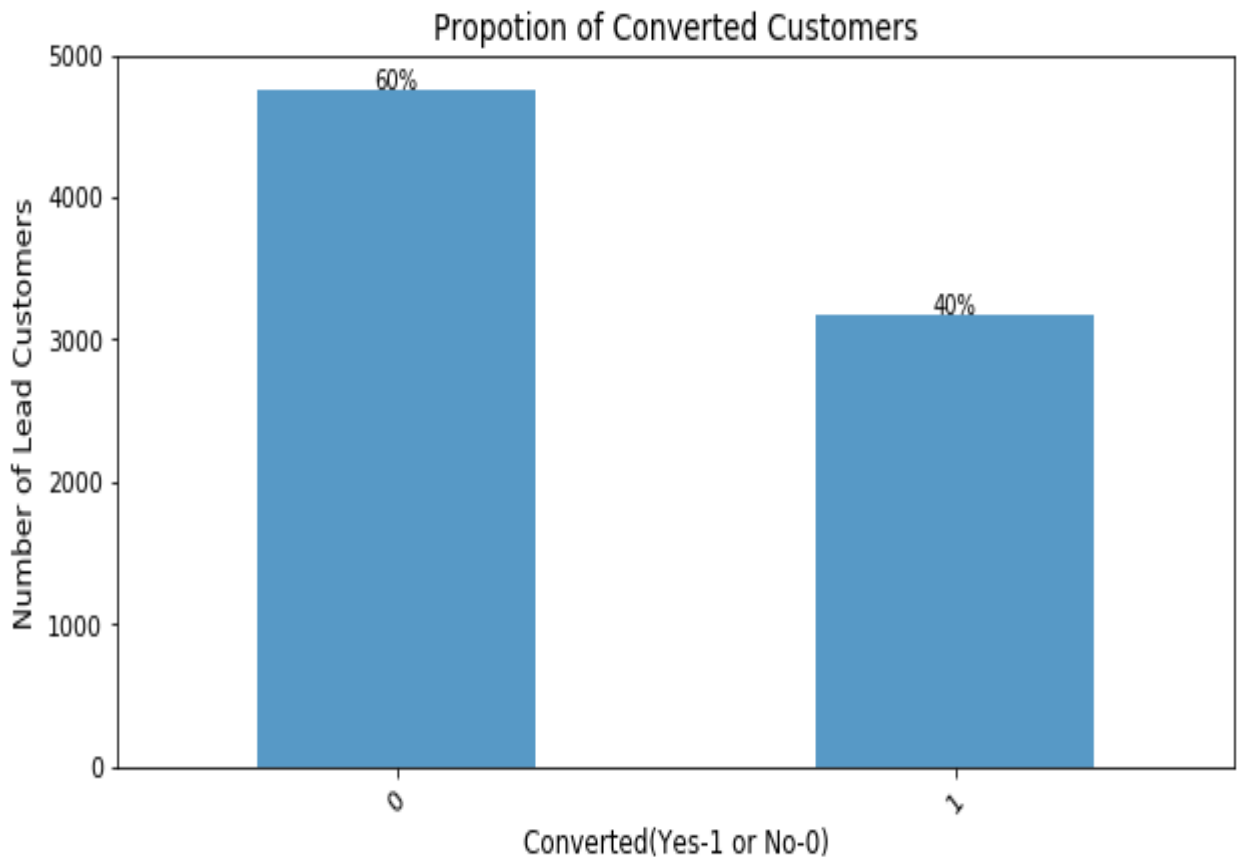


Data Analysis

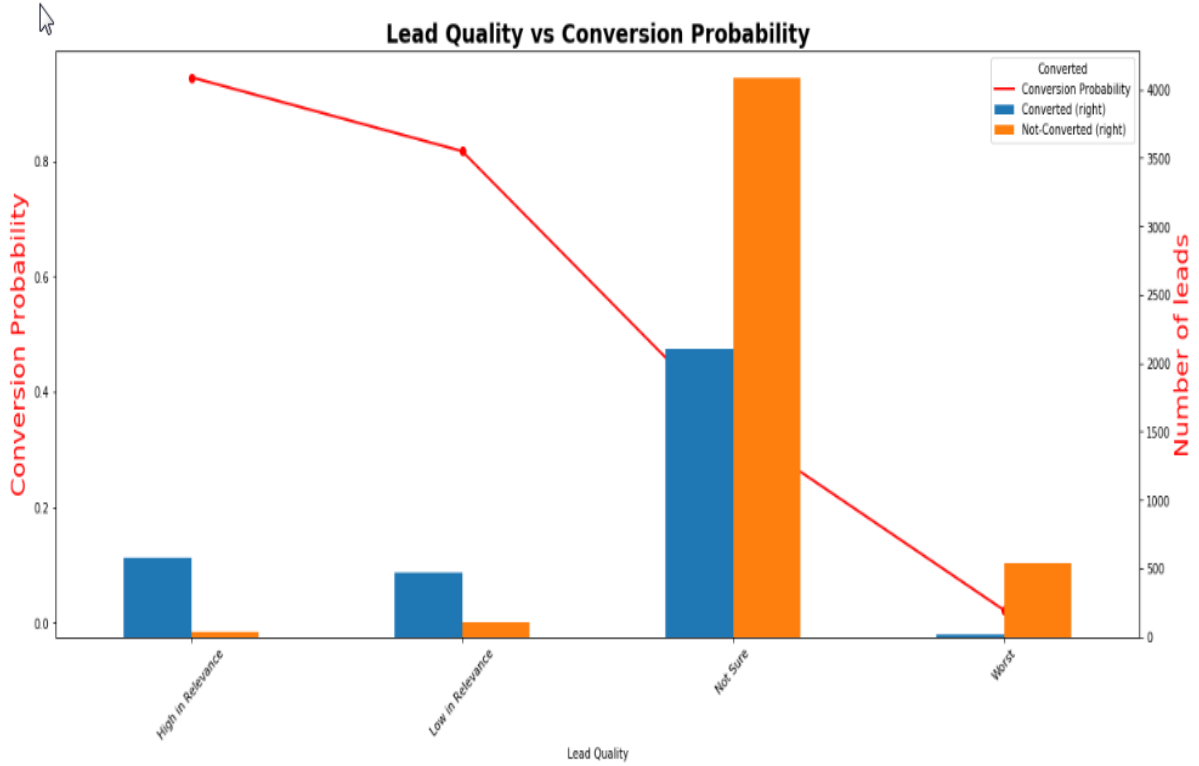
BASED ON TARGET VARIABLES - VISUALIZATIONS

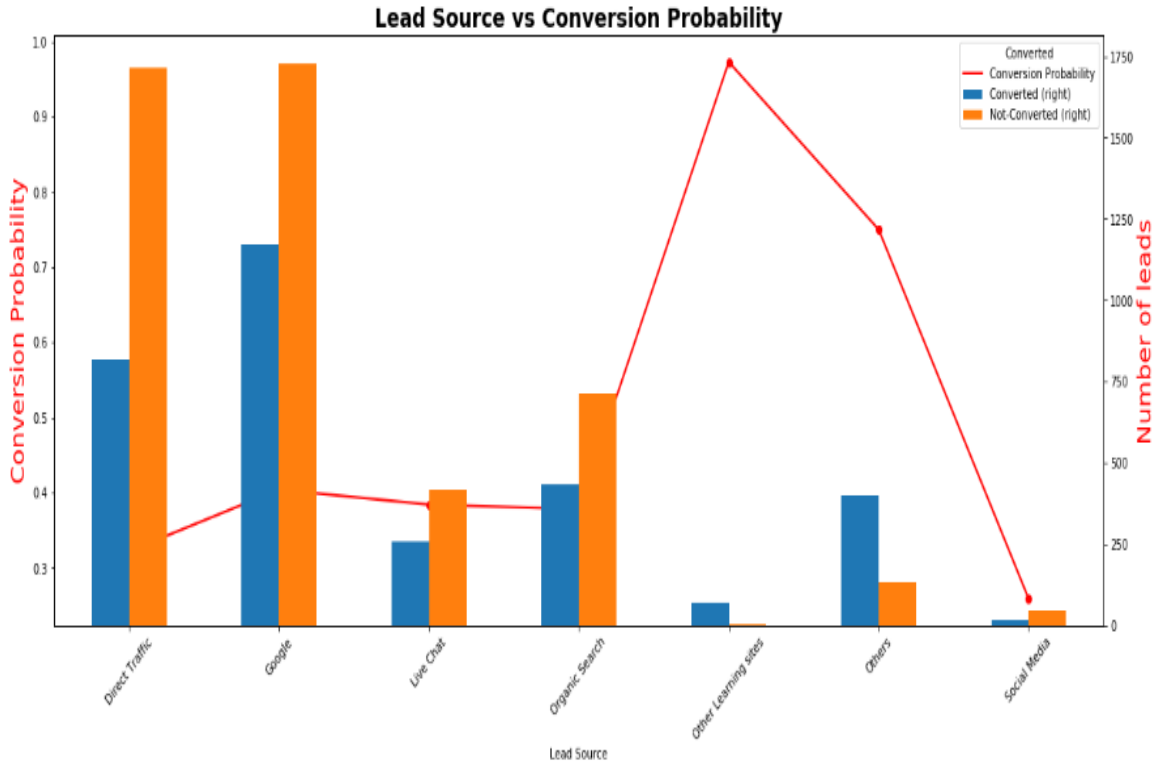
Proportion of Converted Customers

- The number of customers that have converted is less than what the client is expecting.
  - Percentage of Converted Leads in our dataset: approx. 40 %
  - Percentage of Not Converted Leads in our dataset: 60 %



Analysis - Based on Lead Quality and Lead Source

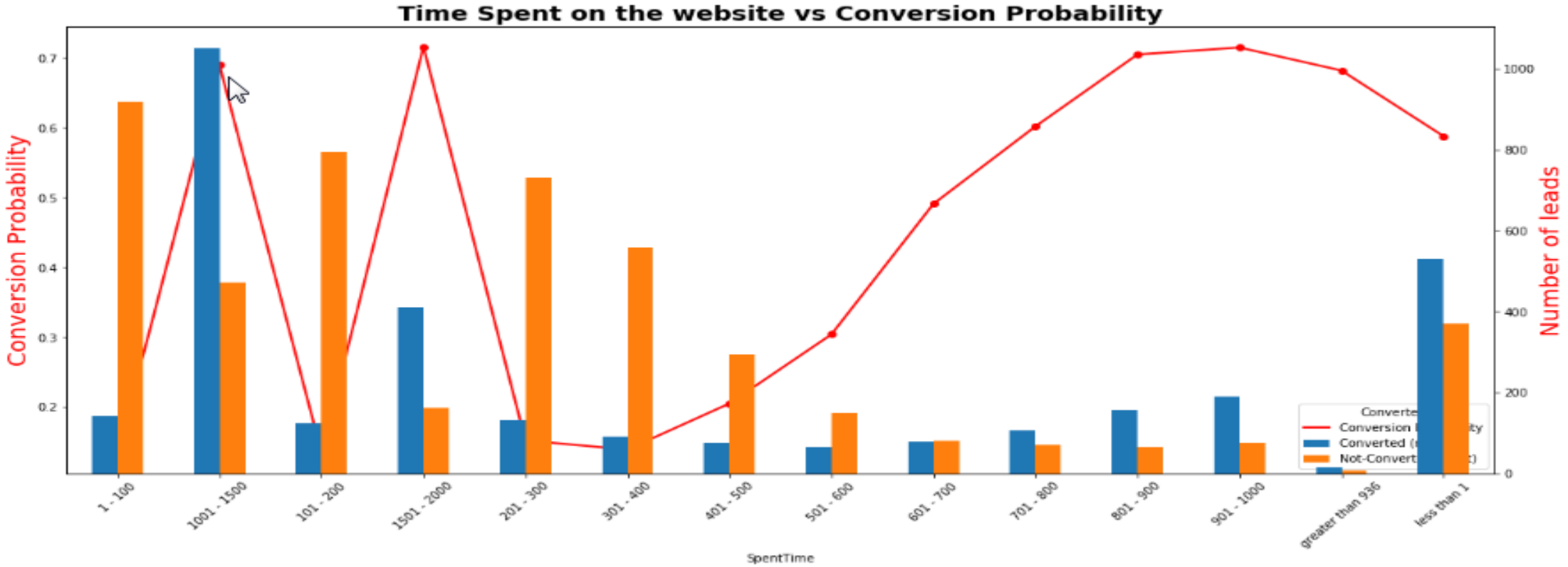




From the above plots, we can infer that,

- ***Customers with High lead quality tend to convert to hot leads and with worst lead quality have a very less priority of becoming a hot customer.***
- ***Conversion Rate is high when the lead source is other learning websites and lead source Social Media has less of conversion rate***

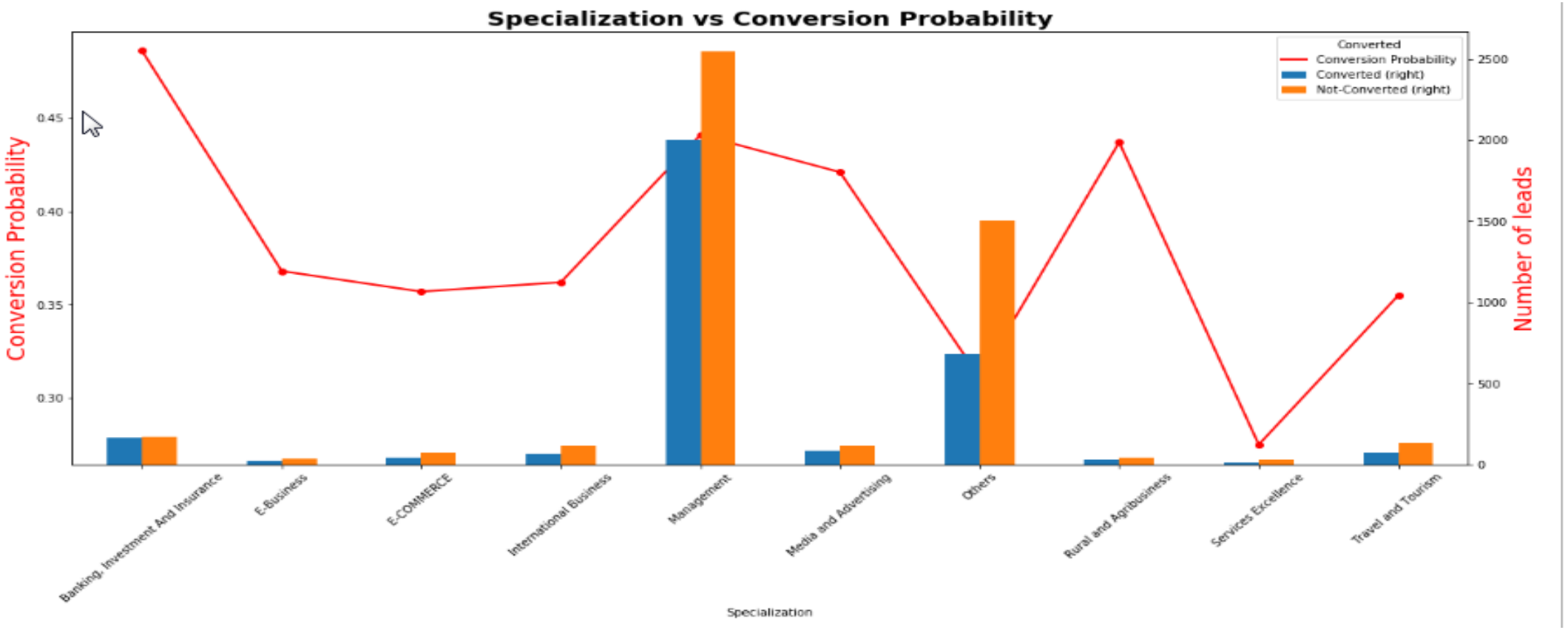
Analysis – Based on Time Spent on Website



From the above plots, we can infer that,

- As expected, the conversion rate is high when the lead spend more hours on the website (more than 1000)

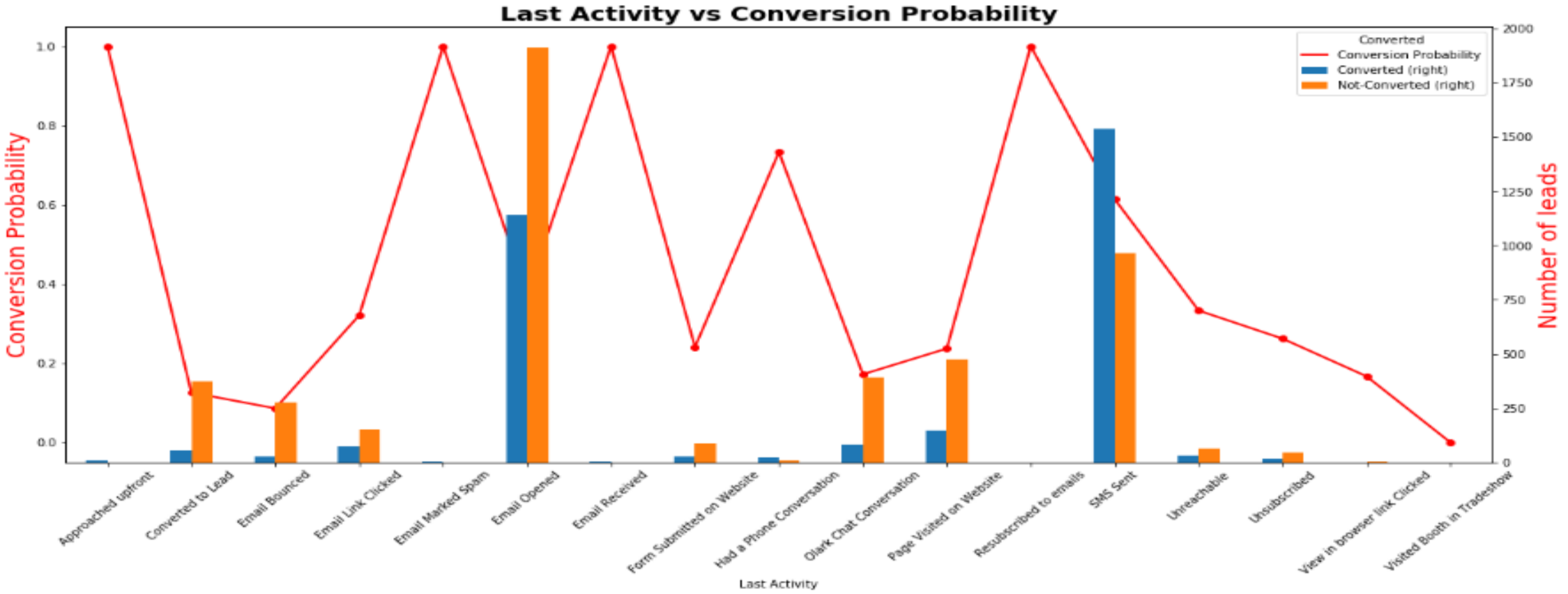
Analysis – Based on Specialization



From the above plots, we can infer that,

- The conversion rate is high when the lead has a specialization in Banking , Investment and Insurance.

Analysis – Based on Last Activity



From the above plots, we can infer that,

- We have highest conversion probability when the Lead Last Activity is related to emails.

Modelling

BUILDING THE MODEL ON THE TRAIN DATASET

PREDICTING THE MODEL USING THE TEST DATASET

Features Selected for building the model on Train dataset

Using RFE , the below features were selected for building the model.

- Lead Source - The source of the lead.
  - Other Learning sites (WeLearn and Welingak Website)
- Tags - Tags assigned to customers indicating the current status of the lead.
  - Closed by Horizon
  - Lost to EINS
  - Ringing
  - Will revert after reading the email
  - switched off
- Lead Quality - Indicates the quality of lead based on the data and intuition the the employee who has been assigned to the lead.
  - Worst
- Last Notable Activity - The last notable activity performed by the lead.
  - Others (Unreachable, Had a Phone Conversation, Email Marked Spam, View in browser link Clicked, Re-subscribed to emails, Form Submitted on Website, Approached upfront, Email Received)
  - SMS Sent
- Total Time Spent on Website

Features selected to Predict the model after training the train dataset

- Tags - Tags assigned to customers indicating the current status of the lead.
  - Closed by Horizon
  - Lost to EINS
  - Ringing
  - Will revert after reading the email
  - switched off
- Lead Quality - Indicates the quality of lead based on the data and intuition the the employee who has been assigned to the lead.
  - Worst
- Last Notable Activity - The last notable activity performed by the lead.
  - SMS Sent
- Total Time Spent on Website

Accuracy , Sensitivity and Specificity for various probabilities on the train dataset

Optimal point seems to be 0.3 for cut-off probability.

ROC Curve using the Converted and Predicted Value

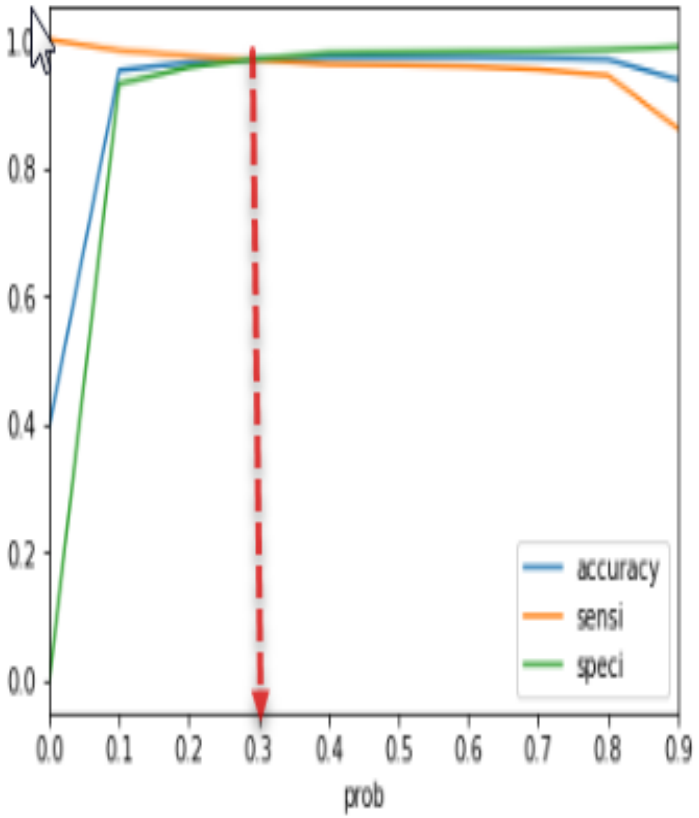
Sensitivity: 0.9673

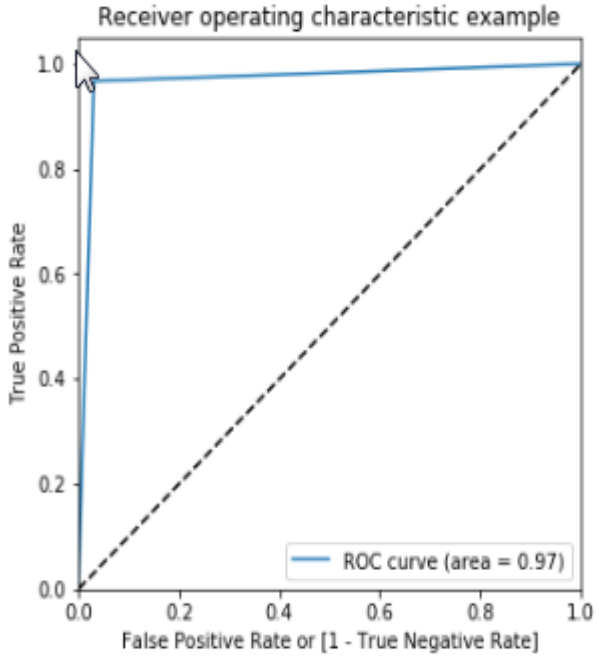
Specificity: 0.9695

Accuracy: 0.9686

In Percentage:

Accuracy, Sensitivity and Specificity are approximately 97%

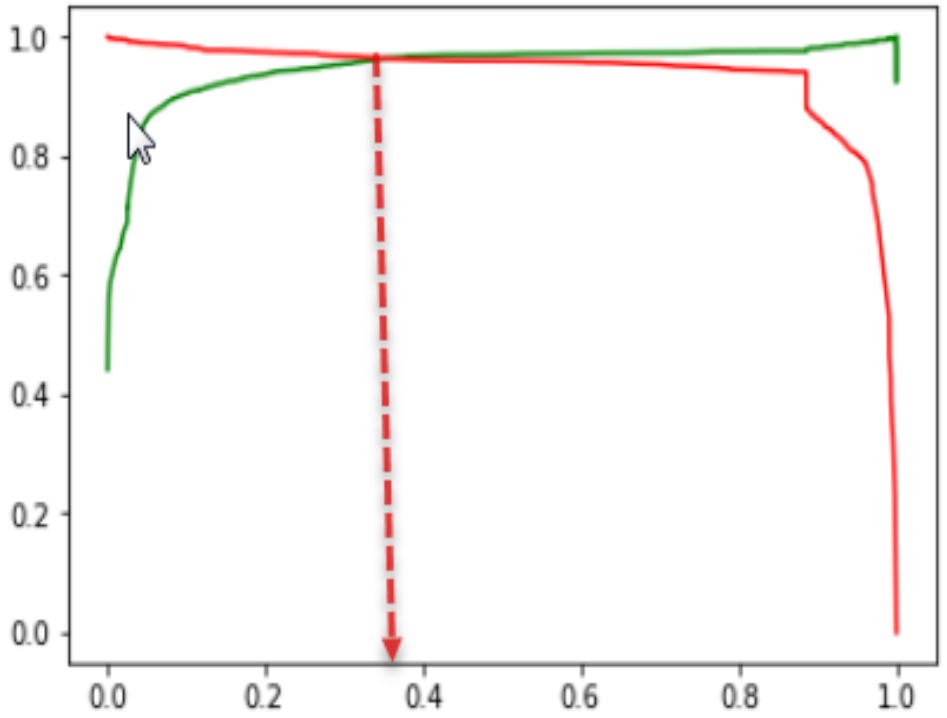




ROC curve area = 0.97

Precision and Recall on the Train data set

- Precision Score : 0.9543
- Recall Score : 0.9673
- The precision and Recall Score is approx. 96%
- From the plot,
- Optimum Precision-Recall cut-off is 0.35



Model Prediction

Based on the Conversion probability which we predicted from the specificity/sensitivity graph of train dataset, we can predict the values in the test dataset.

ROC curve after predicting the test dataset.

Accuracy, Sensitivity and Specificity for various probabilities on the test dataset

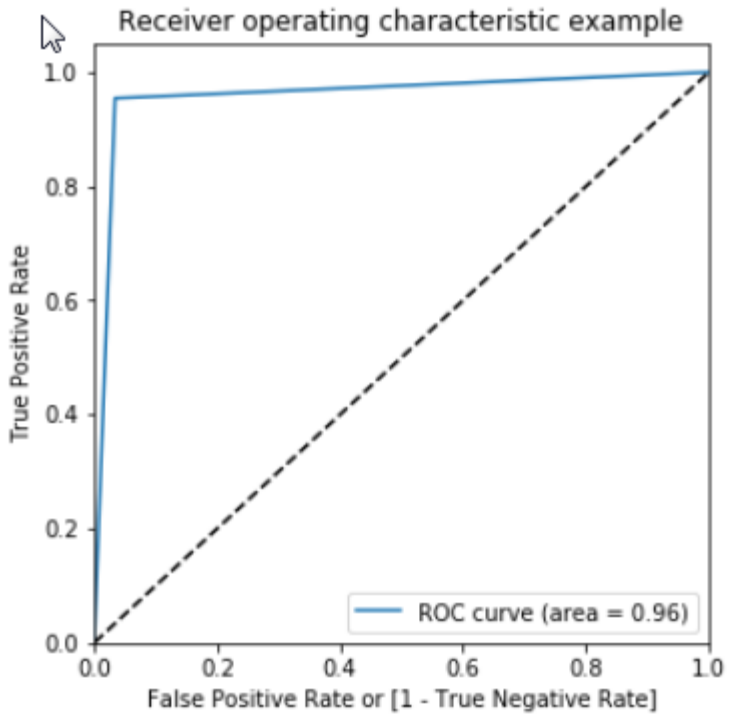
Sensitivity: 0.9545

Specificity: 0.9659

Accuracy: 0.9613

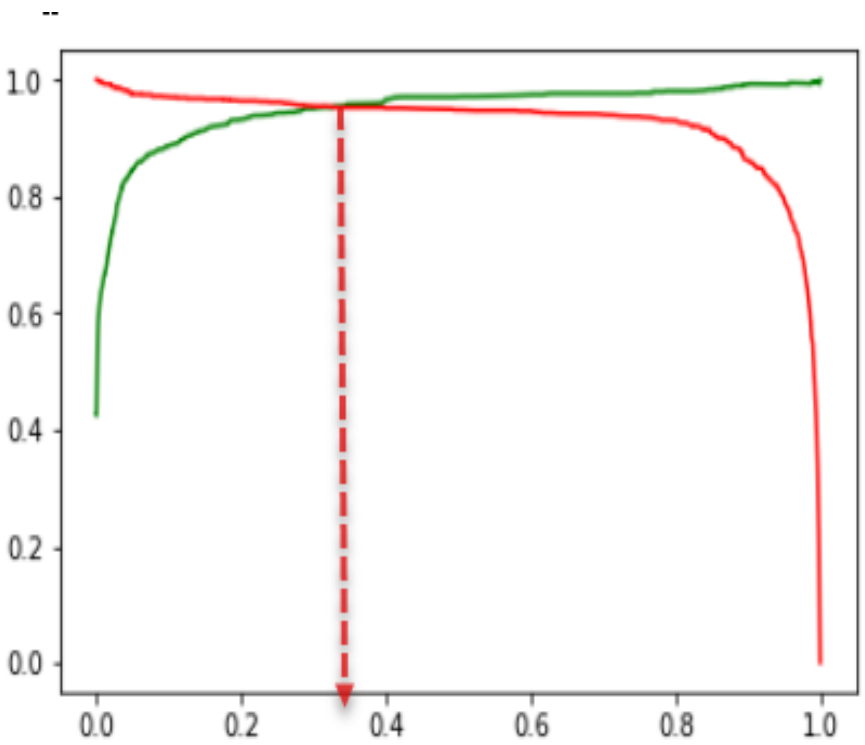
In Percentage:

Accuracy, Sensitivity and Specificity are approximately 96%



Precision and Recall on the Test data set

- Precision Score : 0.9506
- Recall Score : 0.9545
- The precision and Recall Score is approx. 95%
- From the plot,
- Optimum Precision-Recall cut-off is 0.35



Conversion Ratio

After combining the Train and Test datasets and based on the cut off generated from Precision Recall Curve, the below ratio were retrieved.

- Conversion Ratio for cut off probability 0.35 : 96%
- Conversion Ratio for cut off probability 0.2: 94%
- Conversion Ratio for cut off probability 0.1: 96%

The Client is looking for a conversion ratio of approximately 80%

- Conversion Ratio for cut off probability 5% : 85%

We are considering a final cut off probability of 0.05 for getting the desired conversion ratio.

Inferences

BASED ON BUILT MODEL

Key Insights

Top three variables in our model which contribute most towards the probability of a lead getting converted are :

- Tags (Specifically: Closed by Horizzon, Lost to EINS, Will revert after reading the email)
- Total Time Spent on Website
- Last Notable Activity (Specifically: SMS sent)

Top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion are :

- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Tags_Will revert after reading the email.

Strategy to Adopt Based on Resource Availability

Scenario: In case, we have extra resources and company want to convert as much as possible potential leads, they should follow below strategy:

To make sure of conversion of all potential leads, first we need to find all potential leads. This can be done by altering (lowering down) lead score cut-off.

Once we find our potential leads, Company should follow below steps which can ensure their conversion into hot leads:

- The company should provide a Call Back Request Option for the user.
- Make more than one call (at least 5 calls because anything more than that clearly shows that the lead is not interested).
- If the lead calls to get more information, the wait time should be as less as possible.
- Timely demo should be provided to such customers.

Strategy to Adopt Based on Resource Availability

Scenario: In case, we have less resources or company want sales team to focus on new work as well along with handling customers, they should follow below strategy to focus on extremely important leads:

- **Increase Cut-off to increase Conversion Ratio:**

  In the model, we saw that if we increase probability cut-off, conversion rate will also be increased. This will increase our chances to find out hot leads and we can utilize our time and efforts for those customers.

- **Send Optimal Emails:**

  By now, it has been clear that leads who signed up for emails have higher conversion rate. So, we can even create a subset of those customers chosen by first step, to select only those customers who have opted for email services

Recommendations:

Here are the suggestions company can utilise for increasing the Lead Conversion Ratio :

- To make sure of conversion of all potential leads, first we need to find all potential leads. This can be done by altering (lowering down) lead score cut-off.
  - Once we find our potential leads, Company should follow below steps which can ensure their conversion into hot leads:
    - The company should provide a Call Back Request Option for the user.
    - Make more than one call (at least 5 calls because anything more than that clearly shows that the lead is not interested). Follow up calls to a qualified lead. The more we delay to reach out to a qualified lead, the lower the chances of conversion.
    - If the lead calls to get more information, the wait time should be as less as possible.
    - Timely demo should be provided to such customers.

- The users should be able to access information faster in a website. So keeping a check on the performance of the website will also affect the conversion rate.

- Company should add an option to verify Phone number and e-mail addresses of all customers when they fill up the form. Because non-interested customers may give us their dead sim numbers or wrong phone numbers. This way we will be able to eliminate wasting our efforts behind such customers.

- Increase the standards of the Lead Quality. Re-checking the model's lead conversion ratio and lead score on a routine basis.

- Increase Cut-off to increase Conversion Ratio**:** In the model, we saw that if we increase probability cut-off, conversion rate will also be increased. This will increase our chances to find out hot leads and we can utilize our time and efforts for those customers.

- Send Optimal Emails: From the dataset the , it is clear that the Lead conversion rate is higher for the customers who has selected yes to receive email from the company.
  - We can even create a subset of those customers chosen by first step, to select only those customers who have opted for email services. So sending Optimal email is a very good way of increasing the converting the lead to a potential lead.