Lead Scoring using Logistics Regression

Niraj Kumar & Mahiboob Ali
Data Science Program - Oct 2022

Lead Scoring Assignment

Table of Content

- Problem Statement
- Approach
- Observations and Recommendation

Problem Statement

Scenario

An education company X Education sells online courses to industry professionals. It markets itself on various channels and once a prospect fills certain data on its website, its sales agents call leads and try to get them enrolled for a suitable course.

This process involves calling all leads while the **lead conversion is very low – about 30%** Though the investment is high, the returns are low

X Education is unable to **identify Hot Leads** that can be part of a focussed target group to call instead of making calls to all prospects.

Given Data

1. Leads.csv' contains all the information of the lead including prospect id. Other sourcing and communication of various stages in the lead lifecycle are also present. Target variable '**Converted**' shows actual conversion result

2. Leads data dictionary.xls is the data dictionary of the Leads file explain the meaning of each column

3. Assignment Subjective Questions.docx' is a questionnaire of 4 questions asking some specific questions on the model and the strategies to be developed using the model's learnings.

Expected Outcome

Identify Hot Leads by assigning them a lead score between 0 to 100.

- Analyse data, perform EDA
- Build a Logistics Regression model to find what features affect the Converted flag most
- Select a model with high sensitivity
- Identify key features

Business benefit -
This will ensure that when a lead is generated, and during its lifetime, its lead score is updated so that sales team can call Hot Leads and convert them to paying customers

Approach

- ✓ Read the *Leads data dictionary’* file to build and understanding of the data and the domain. This will make us familiar to the data set and build understanding of the business. For ‘*Leads*’ data set do as below

- ✓ Load the data and check the information to see its meta data and column formats

- ✓ Do EDA –
  - ✓ There are many cells that have value ‘Select’. This comes from drop down menus that have a default value ‘Select’. These columns are optional and many times, user does not pick from any available option. Hence ‘Select’ becomes the filled in value. Therefore ‘Select’ must be imputed with null
  - ✓ Any column having more than 45% null values can be considered to be not useful and hence deleted
  - ✓ There could be columns that have < 45% null values. For such categorical columns, there are two ways to impute values
    - ✓ Mode can be imputed
    - ✓ Low frequency (<2%) column values can be grouped into another label and that could be imputed
  - ✓ Rows where some columns are empty and cannot be imputed can be deleted if such rows are less than 2%
  - ✓ Imbalance – Delete columns that have more than 90% imbalanced data
  - ✓ Outliers - Find outliers. They can be left if spread uniformly outside boundaries or such records can be deleted if very far off from adjacent values

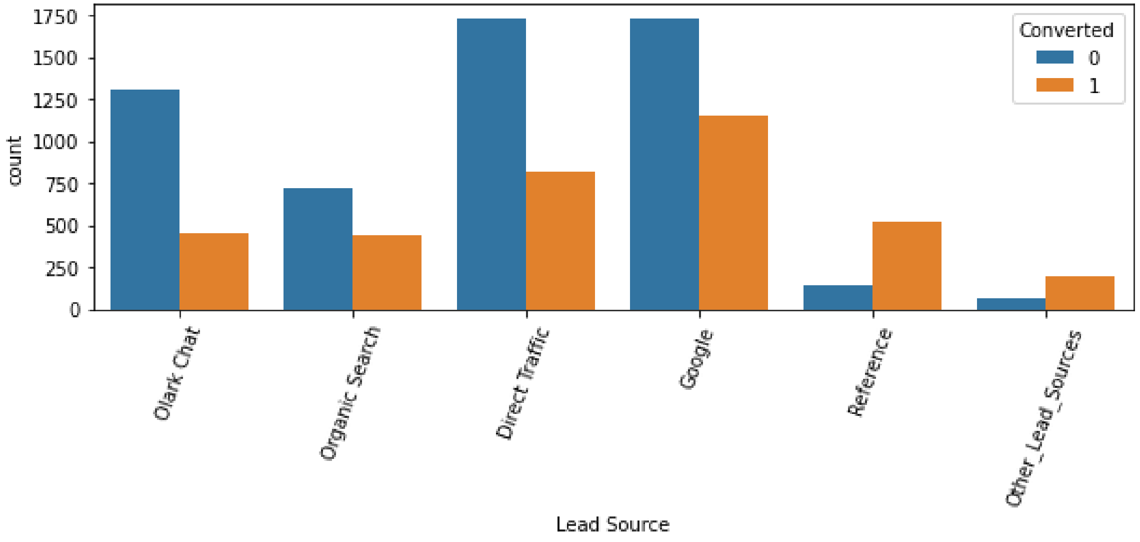
- ✓ Perform Univariate Analysis – Chose sets of parameters and plot them to see patterns

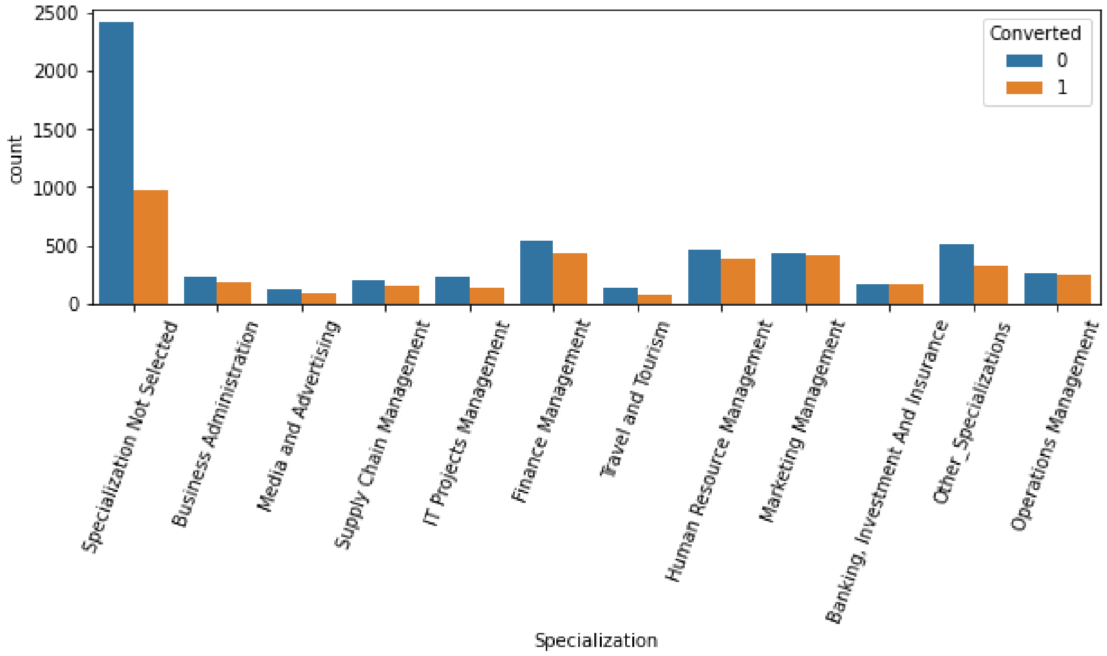
- ✓ Created Dummy variables for categorical columns ignoring one column and delete the column for which dummy variable is created

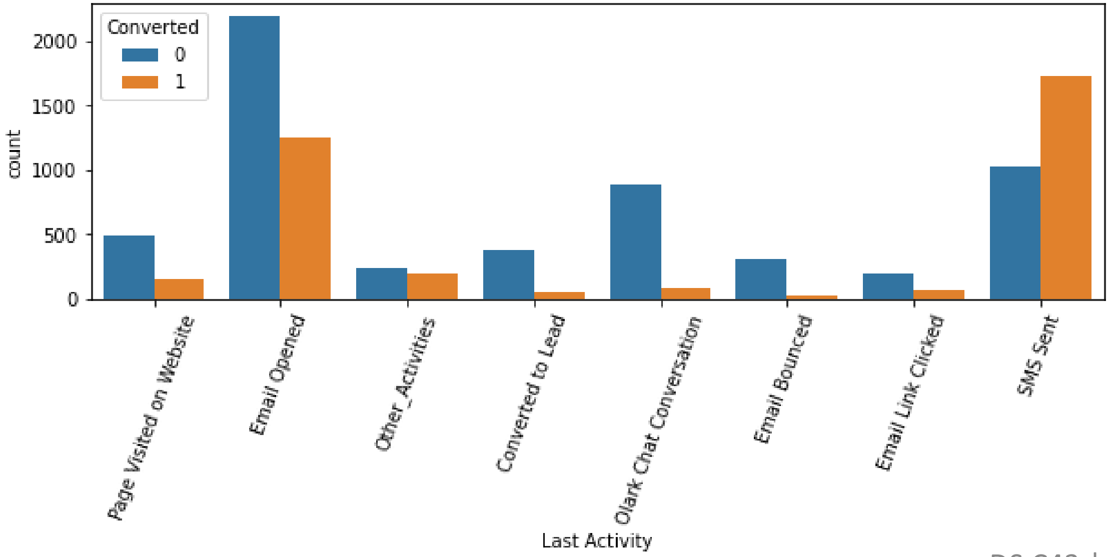
Approach

- ✓ Split the data into 70% training and 30% test sets with some randomness value
- ✓ Scale the training data using standard scaler
- ✓ Build first model using all features
- ✓ Using RFE select 16 features. Calculate accuracy, Sensitivity and Specificity. See VIF score. Reduce features one by one based on p values or VIF scores. Then build new model and calculate scores again. Do this iteratively till the time number of columns is under 15 and the model does not improve further.
- ✓ Calculate ROC curve to see model performance.
- ✓ Scale test data and apply model to test data. Calculate performance measures from test data
- ✓ Perform Principal Component Analysis to see the features that contribute maximum to the model

Observations and Recommendations



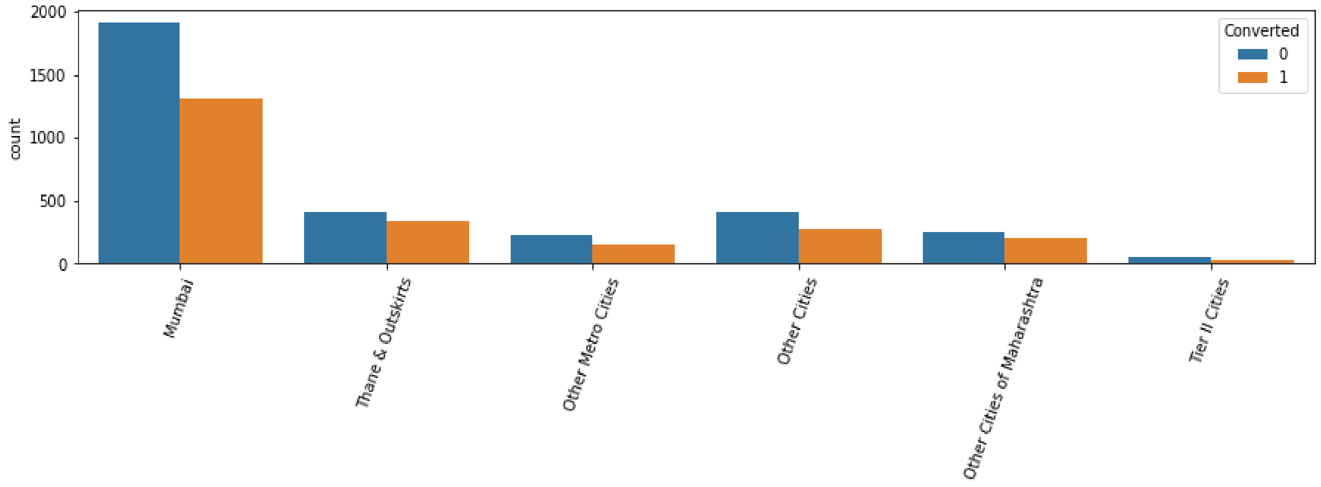


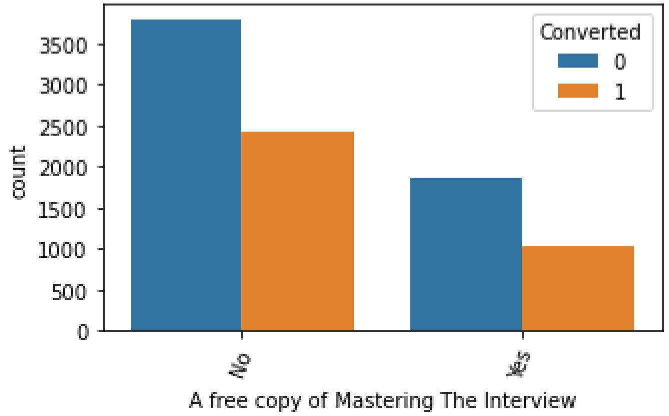


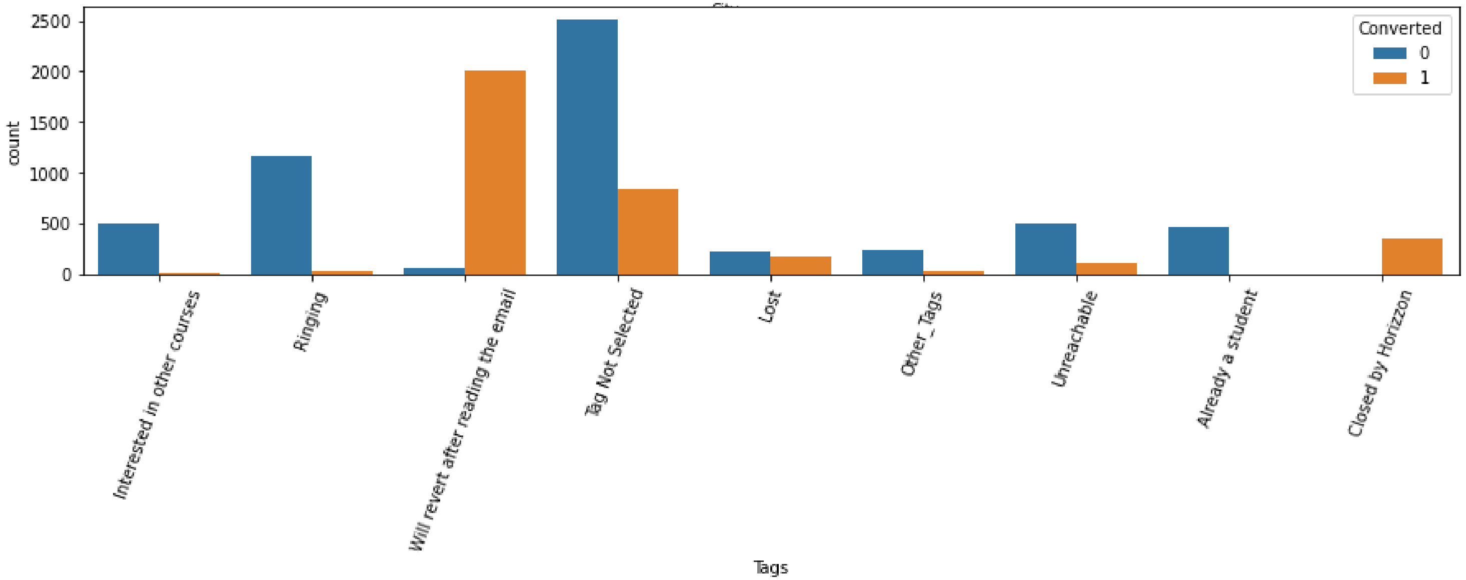
Low Frequency data is grouped together in categorical columns

- Direct Traffic, Google, Reference and Olak Chat are best sources of Leads
- If lead is exchanging SMS and chat then they have high chance of conversion
- Most of leads do not disclose specialization

Observations and Recommendations



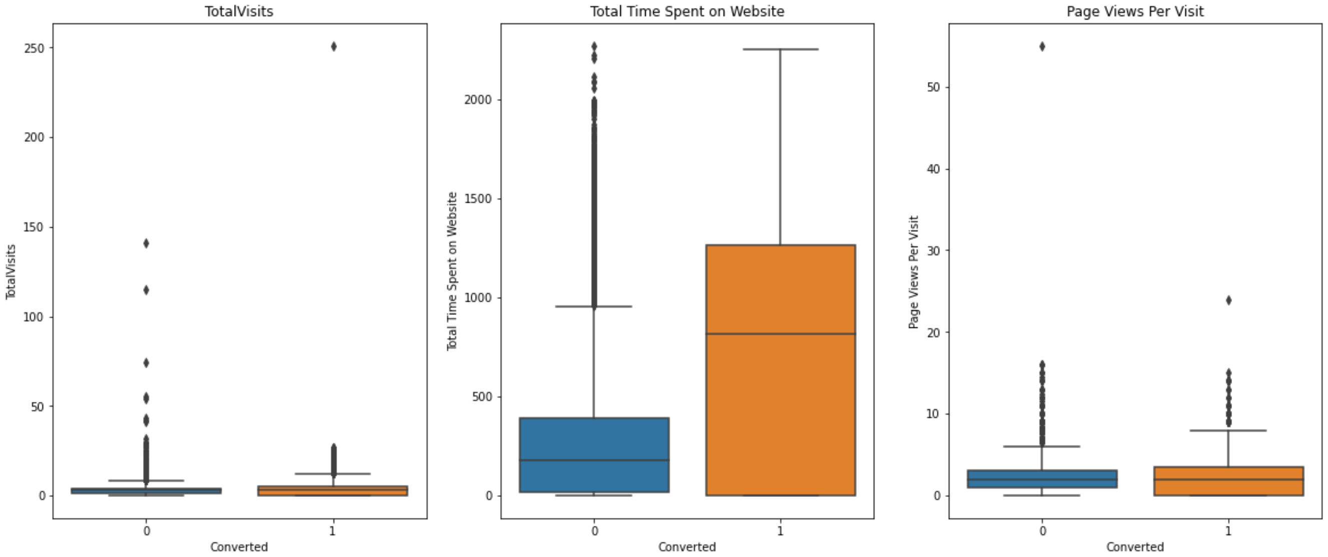


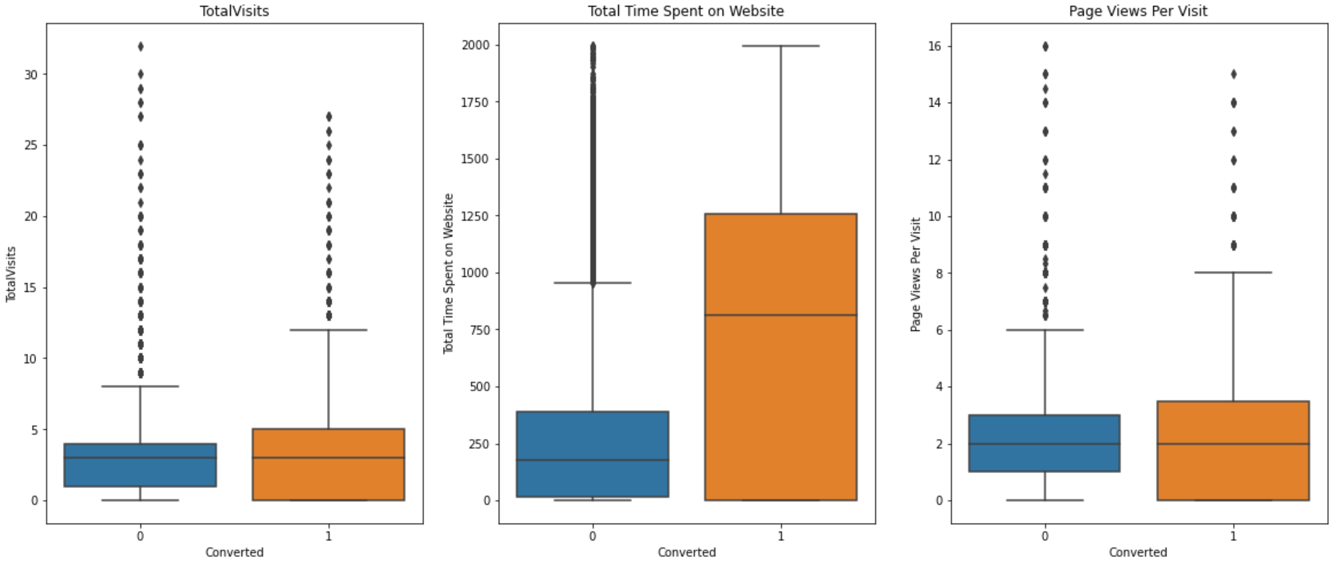


Low Frequency data is grouped together in categorical columns

- City Wise Mumbai and Thane are geographically adjacent and show same behaviour, so they can be combined. Then the data becomes imbalanced and this column can be dropped
- Getting a copy of Mastering the Interview is not affecting lead behaviour and eventual conversion
- Based on lead actions, if lead has committed to sending email, then they have good chance of conversion

Observations and Recommendations

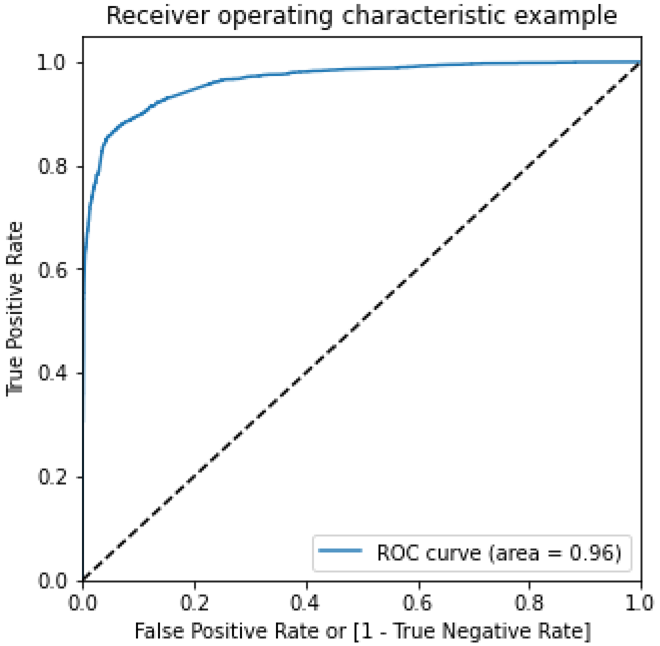




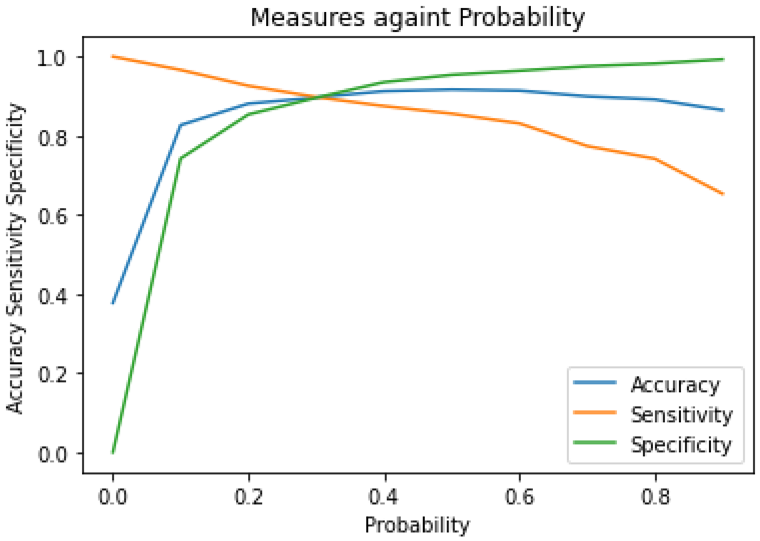
Numerical data – Total Visits, Total Time spent on website and Page Views per Visit

- All of these have outliers. Only some of the records are removed so that the spread across upper limit is uniform
- We can see that if person is spending more time on website then there is a good conversion ratio. This means that the person knows what he/she wants and looking for focussed information. If such people can be targeted well by sales team, then sales outcome will be good

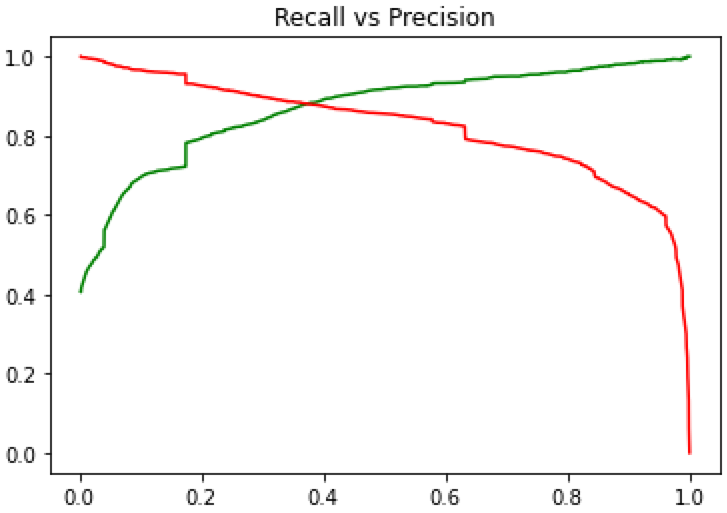
Observations and Recommendations



ROC Curve - The curve follows the left-hand border closely and then the top border of the ROC space therefore the test is more accurate and the model built is good

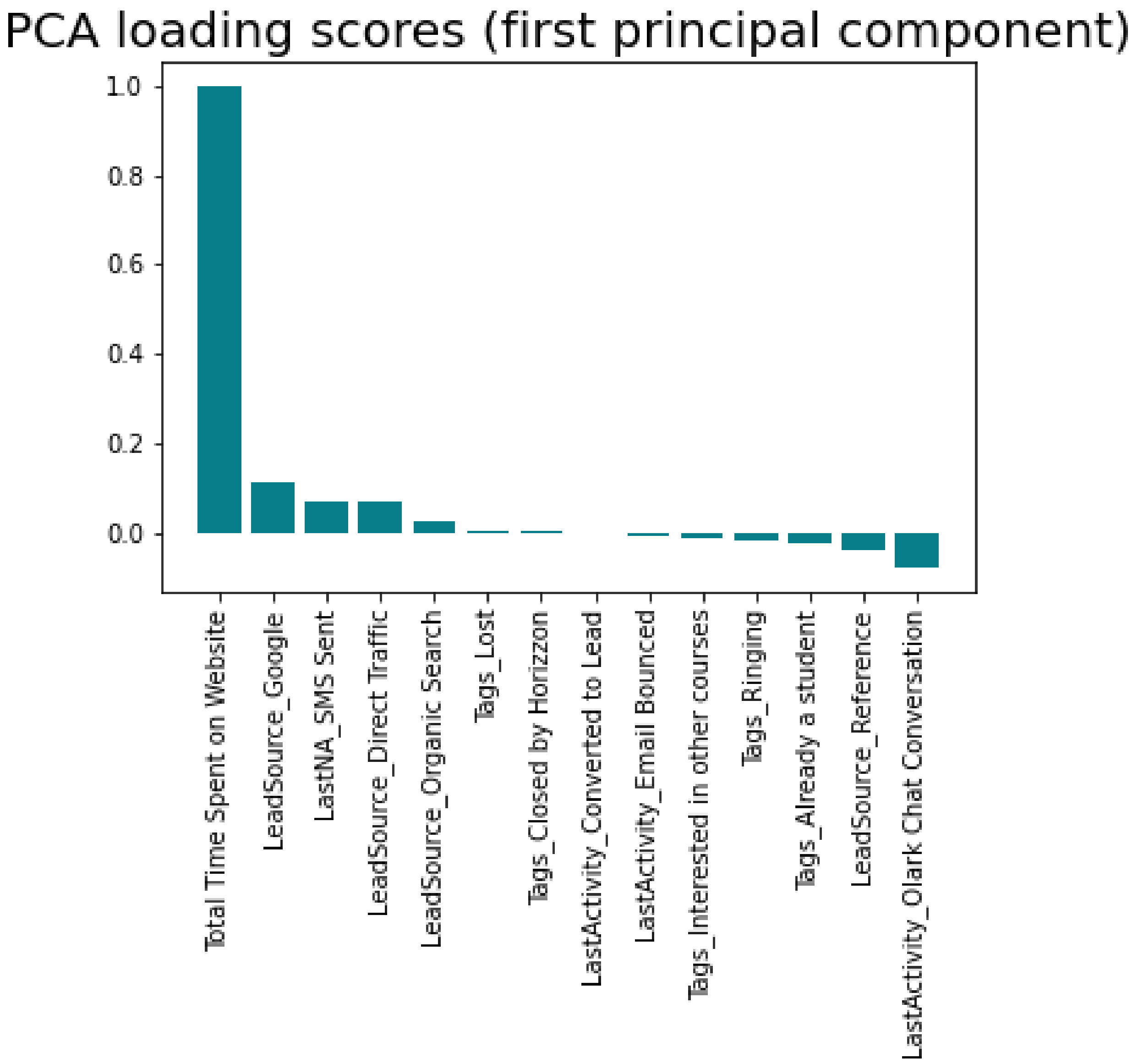


Accuracy Sensitivity and Specificity converge at close to 0.4 and that should be the optimum cut off



Recall and Precision converge at close to 0.4 and that should be the optimum cut off

Observations and Recommendations


PCA loading scores (first principal component)

X Education should target prospects

- That spend high amount of time on their website
- Whose leads have come from Google or Organic search methods
- That engage in communication exchange like SMS

These are people who are motivated as they are doing some research on their own. So if the right information is shared via regular engagement then such prospects can be converted to customers.

X Education can also focus on non call based communication strategy such as email and SMS. The recipients can still be leads that are relatively less hot but can be engaged for a longer period of time using such communication channels. The best is if these channels can be automated then their team can focus on other activities as well