Lead Score case study

Presented By,

Akarsha Hegde

Saket Prusty

Shanthi Reddy

Agenda

- The purpose is to optimize the Lead scoring mechanism based on the leading score with the help of Logistic Regression Model.
- Main Aim is to successfully identify the sets of hot leads

Problem Statement

- An education company named 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.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- In order to increase the Lead conversion rate company should identify most potential leads known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone

Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use

Goals

- 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. 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. 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 so you will need to handle these as well.

Approach

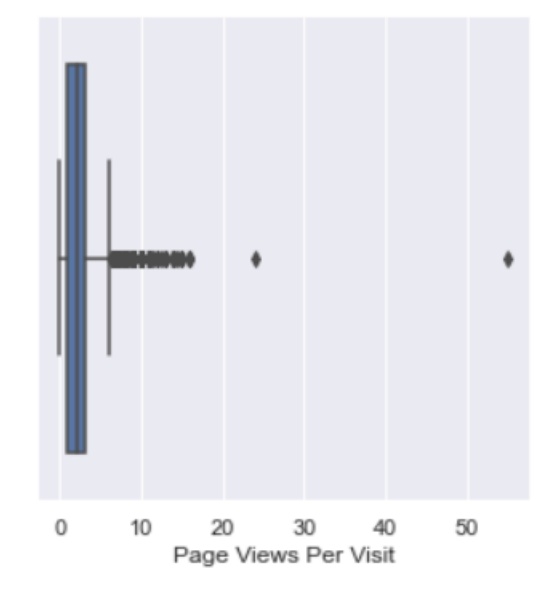
- Reading and understanding the data
- Data Cleaning
- EDA
- Feature Scaling
- Splitting the data into test and train data set
- Prepare the data for modelling
- Model Building and Evaluation
- Making predictions on test set
- Conclusion

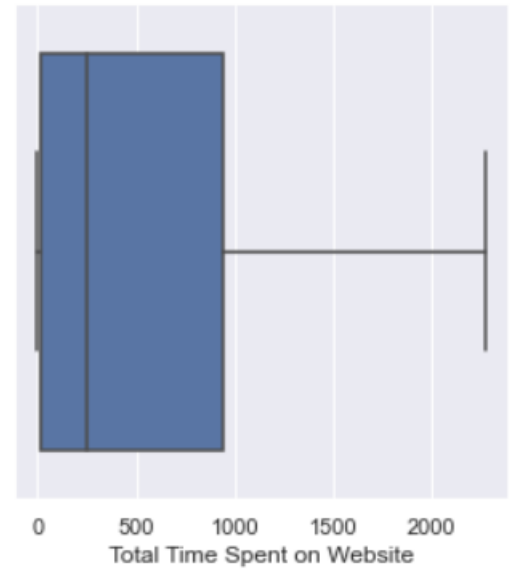
Data Sourcing and cleaning

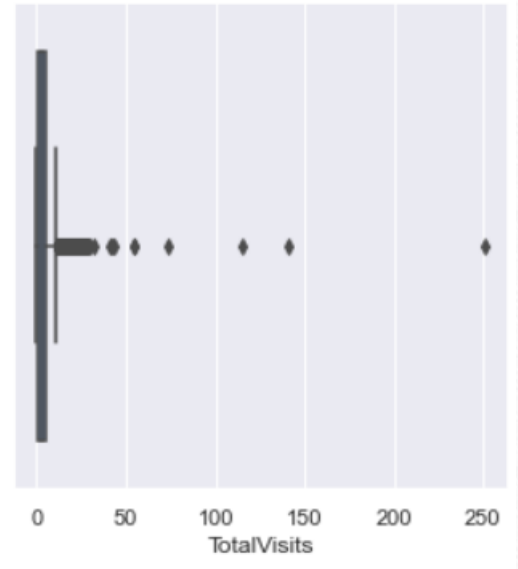
- Read and understand the data from CSV file
- Removing higher null values
- Outlier treatment
- Removal of redundant columns
- Imputing null values
- Exploratory data analysis where conversion rate is 38%
- Feature standardization and Building models
- Analysis based on model built

Outliers

- We can see Outliers for below 3 numerical columns i.e. Total visits, Total time spent on website, Page views per visit have outliers

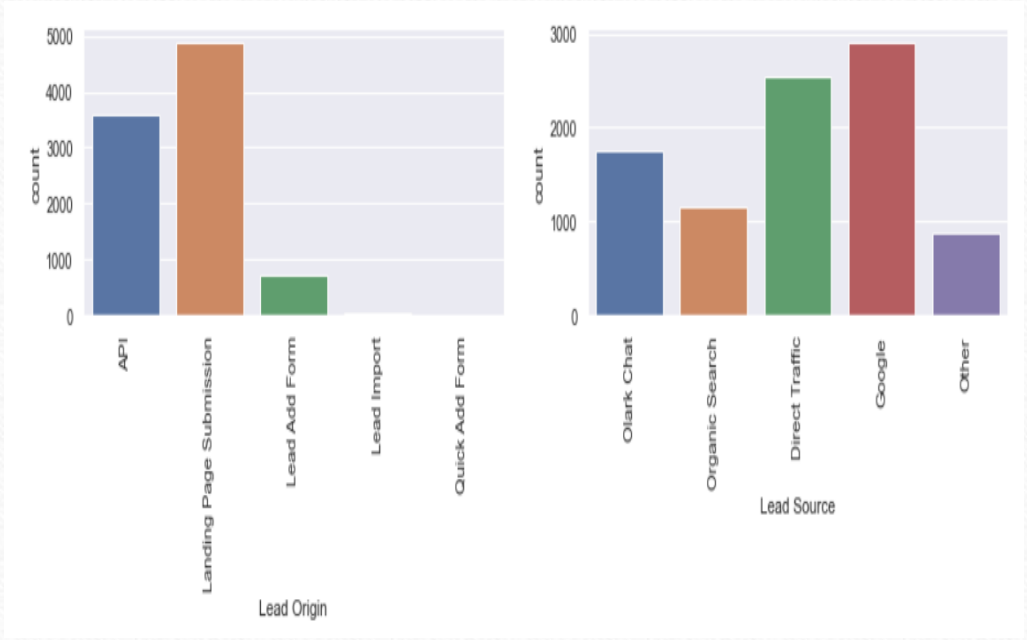


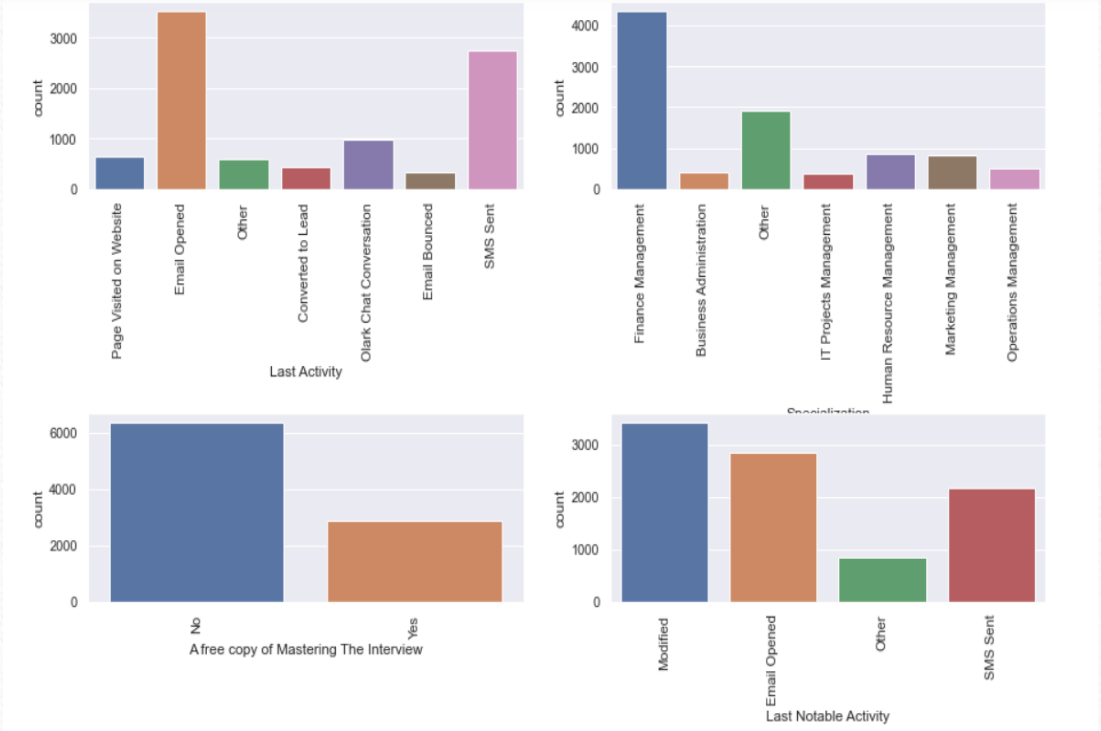




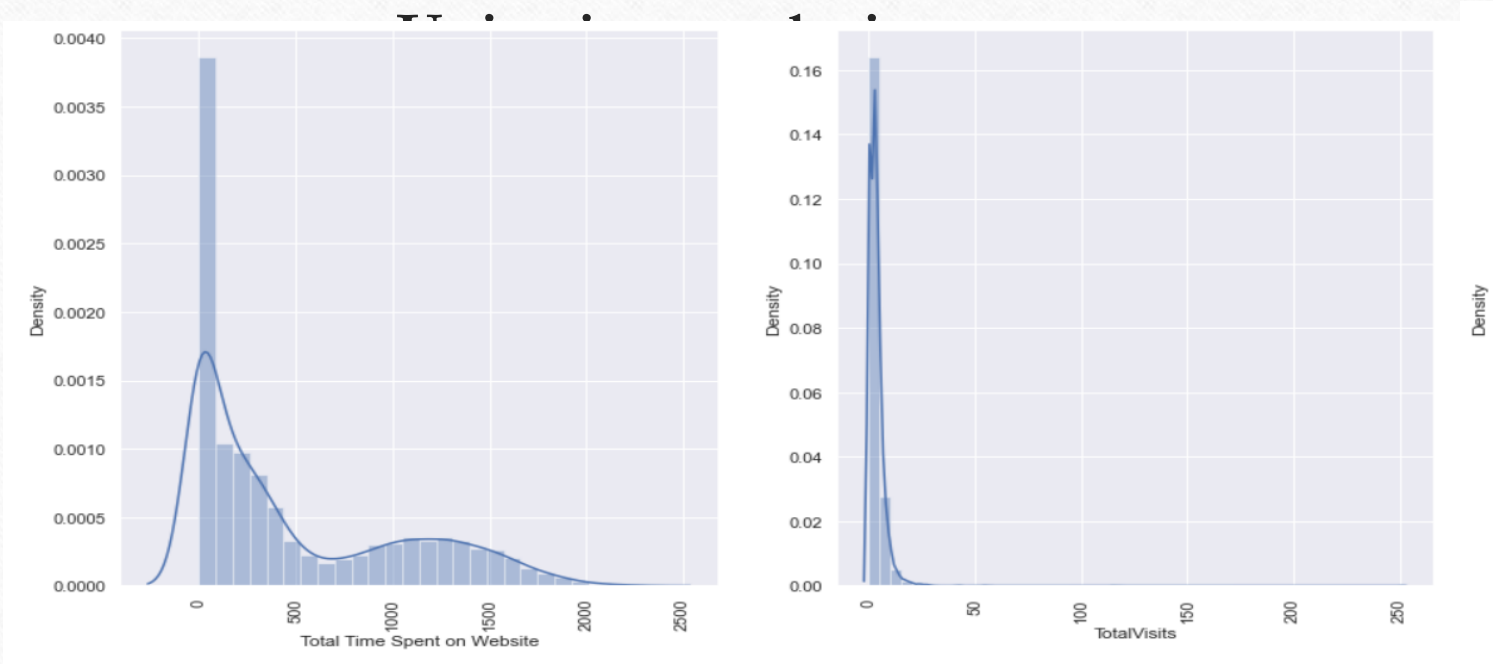
Data Analysis

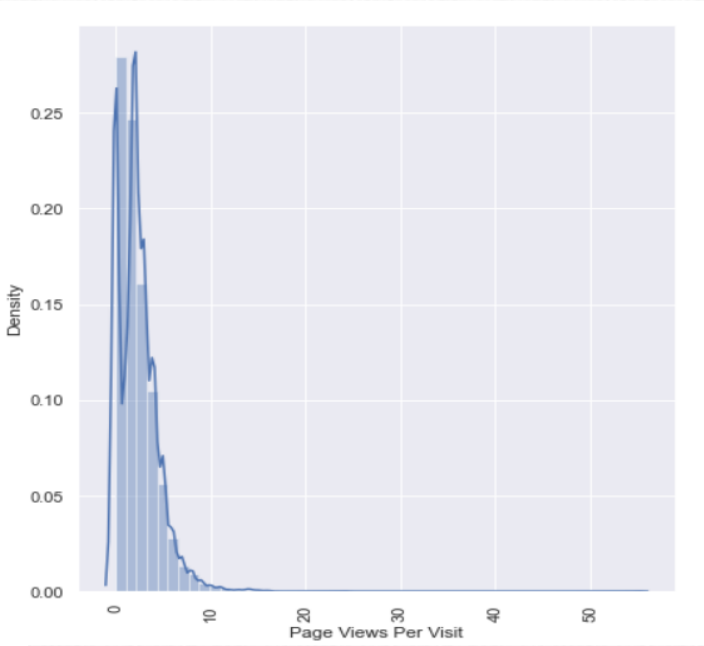
Univariate Analysis on Categorical columns:



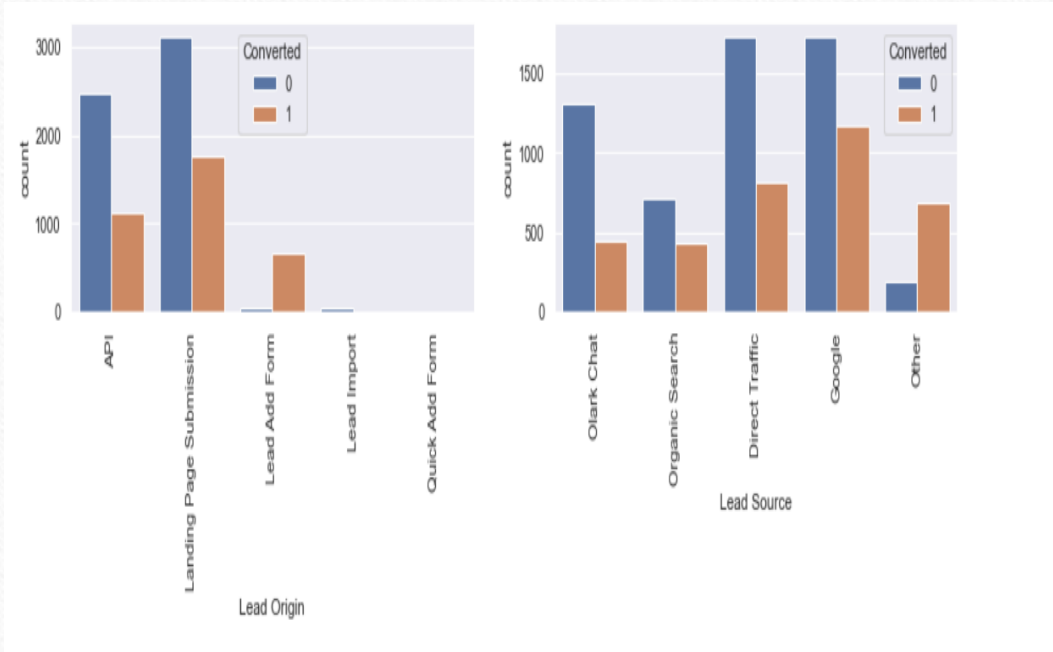


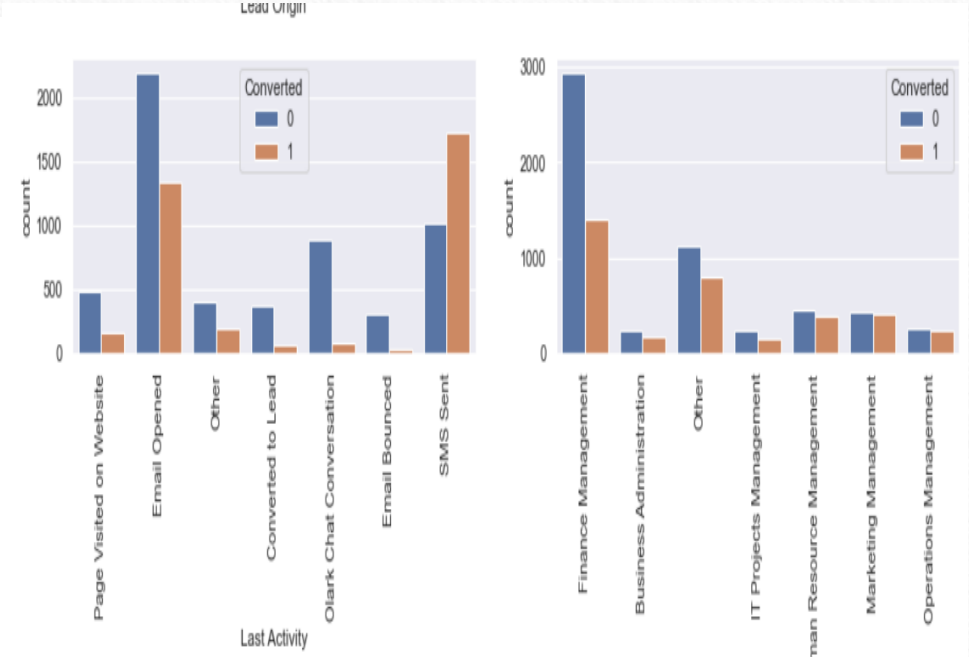
Continuation



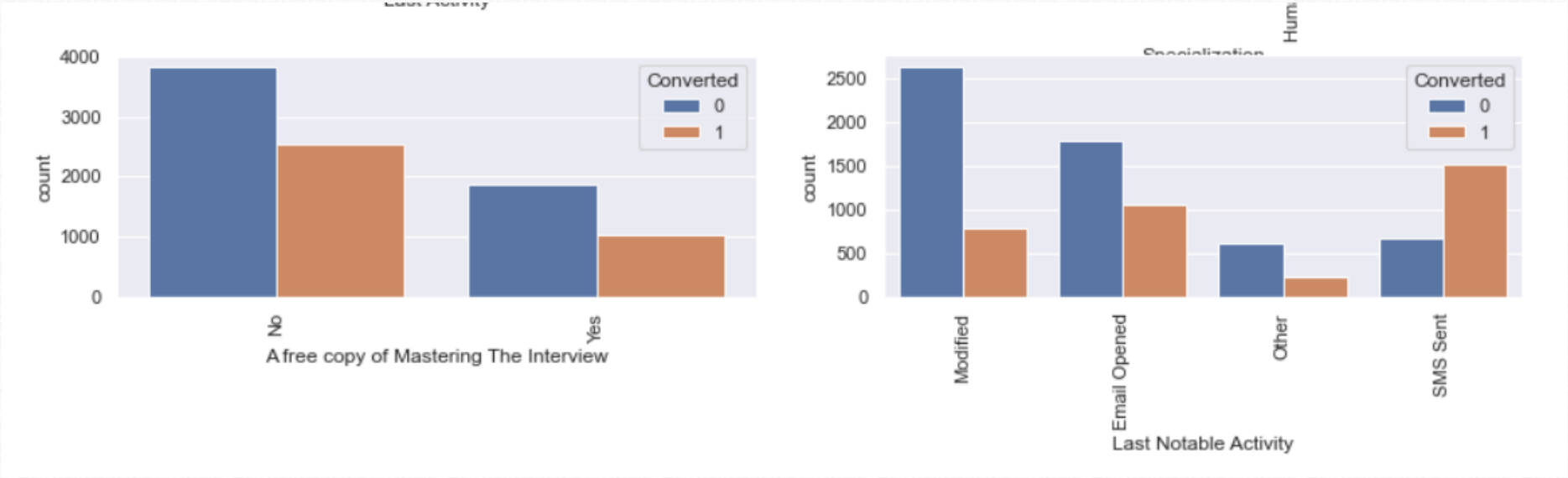


Bivariate Analysis with target as converted





Cont..


4000
3000
2000
1000
0
count
Converted
0
1
No
Yes
A free copy of Mastering The Interview
2500
2000
1500
1000
500
0
count
Converted
0
1
Modified
Email Opened
Other
SMS Sent
Last Notable Activity

Data preparation

Data preparation (also referred to as “data pre-processing”) is the process of transforming raw data so that data scientists and analysts can run it through machine learning algorithms to uncover insights or make predictions.

- Converted yes or no data into binary variables in to 0 and 1
- Created dummy variables

Feature Scaling

Feature scaling is an important pre-processing step in machine learning that helps to ensure that all features are on a similar scale.

- Feature scaling is performed on continuous variables i.e. 'Total Visits' , 'Total Time Spent on Website',' 'Page Views Per Visit'.

Splitting test and train sets

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. It is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for your predictive modelling problem

- Here we have split data in to test and train sets where train size is given as 0.7 and test size is given as 0.3.

Model Building

- Feature Selection using RFE
- Determined optimal model using Logistic regression
- Calculated Accuracy, Sensitivity, Precision and Recall
- Evaluated Model

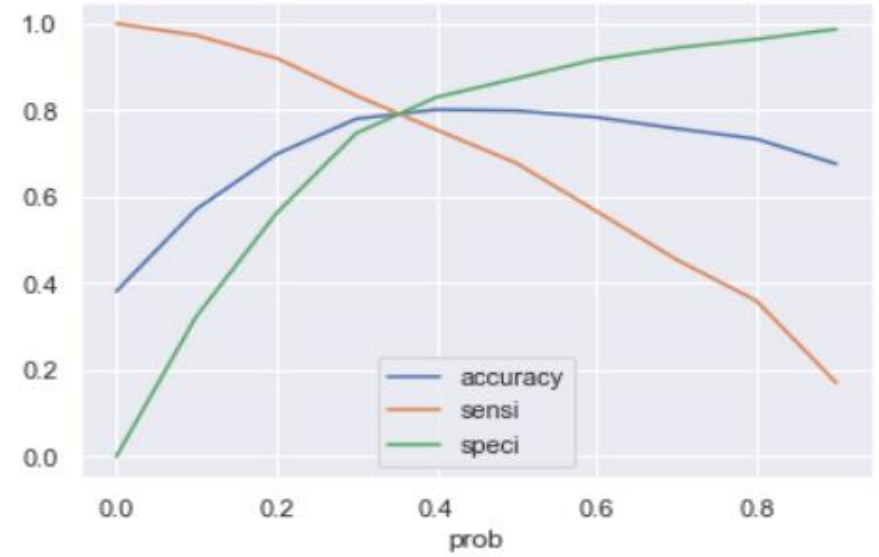
Variables impacting the conversion rate

- Total Visits
- Total time spent on website
- Page Views per visit
- Lead Origin API
- Lead Origin Landing page submission
- Lead Source -direct traffic, google and other
- Last Activity-converted to lead, Email bounced, Olark chat conversion, SMS Sent, Page visited on website, other
- Specialization-Business Administration, Finance Management, Human resource management

Model Evaluation

Accuracy, Sensitivity and specificity on train set

- Accuracy =79.6%
- Sensitivity=77.77%
- Specificity=80.7%

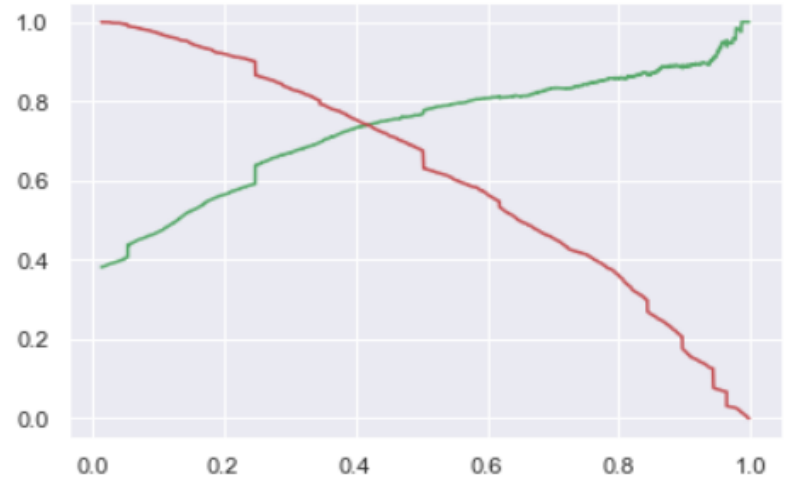


From the above figure we can take 0.37 as cut off

Model Evaluation

Precision and Recall on train data set

- **Precision = 71.34**
- **Recall = 77.7**



The above graph shows the trade-off between the Precision and Recall is 0.41 .

Model Evaluation

Accuracy, Sensitivity and specificity on test set

- Accuracy =78.42%
- Sensitivity=76.80%
- Specificity=79.48%

Result

- Accuracy, Sensitivity and Specificity values of training set are close to test set
- Accuracy, Sensitivity and Specificity values of training set are 79.6%,77.77%,80.7%.
- Accuracy, Sensitivity and Specificity values of test set are 78.42%,76.80%,79.48%
- We have done prediction on the test set using cut off threshold from sensitivity and specificity metrics

Conclusion

- We have checked Sensitivity-specificity and precision-recall metrics and also considered optimal cut off based on sensitivity and specificity for the final prediction.
- Accuracy, Sensitivity and Specificity values of training set are close to test set
- Overall Model looks fine

Summary

There are lot of leads generated in the initial stage but few come out as paying customers in the bottom. In middle stage one has to nurture the potential leads well in order to get higher leads conversion. First sort out best leads prospectus from the leads you have generated. ‘Const’, ‘last activity was SMS Sent’, ‘Lead source Olark Chat,’Total visits’ which contribute more towards leads getting converted. You must keep a list of leads handy so that you can inform them about new courses. Give all necessary information which interest the leads. In this way one can attract more leads.