Lead Score Case Study

To build a Logistic Regression Model to predict whether a lead for online courses for an education company named X Education would be successfully converted or not

Presented By :

V. Prabhakar and Narasimha Reddy N L

Problem statement

□ X Education sells online courses to industry professionals.

□ 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.

□ To make this process more efficient, the company wishes to identify the most potential leads, also 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.

Solution Methodology

Basic steps:

□ Import Data and all libraries, inspect the data frame

□ This helps to give a good idea of the data frames.

Data cleaning:

1. Check and handle duplicate data.
2. Check and handle NA values and missing values.
3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
4. Imputation of the values, if necessary.

EDA:

□ Univariate data analysis: value count, distribution of variable etc.

□ Bivariate data analysis: correlation coefficients and pattern between the variables etc.

□ Feature Scaling & Dummy Variables and encoding of the data.

□ Classification technique: logistic regression used for the model making and prediction.

□ Validation of the model.

□ Model presentation.

□ Conclusions & Recommendation

Cleaning data:

□ The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information.

□ Dropping the columns having more than 40% missing values impute the missing value with mean & mode.

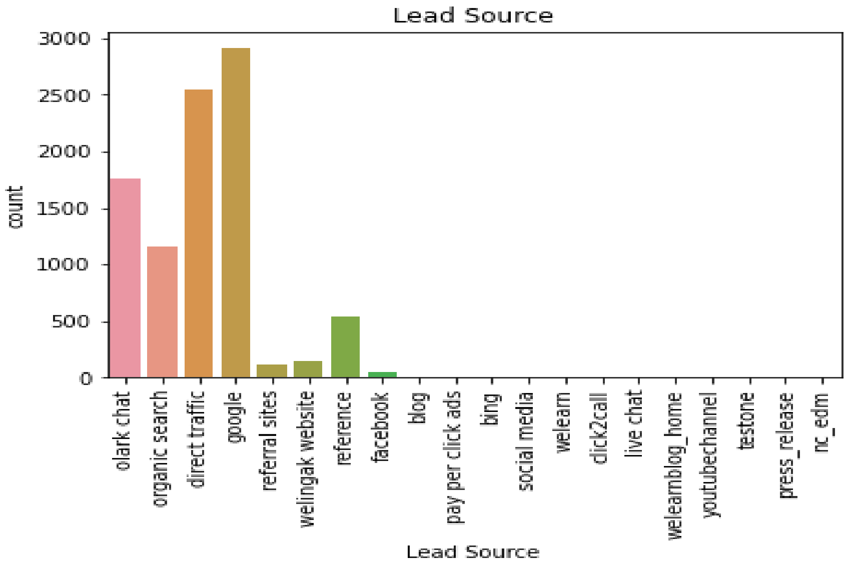
□ Grouping the response which is having very less count.

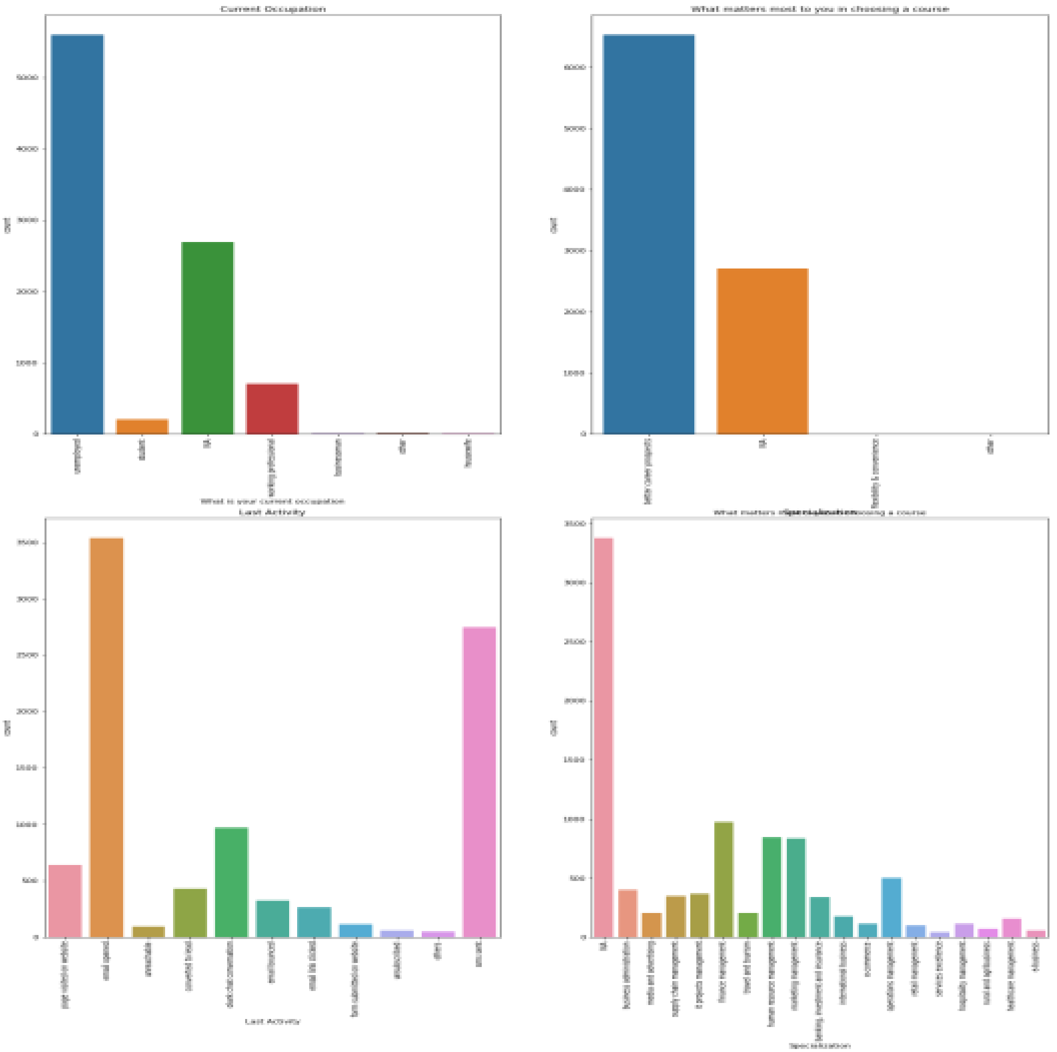
□ Few of the null values were changed to ‘NA’ so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to ‘India’, ’NA’ and ‘Outside India’ child_mort

EDA:

□ A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good and no outliers were found. There are outliers in the numeric variables, Capping all numeric variables of these outliers will help in the analysis

Categorical Analysis





□ from above plot we can see clearly from Lead source and from other variables converting into leads are :

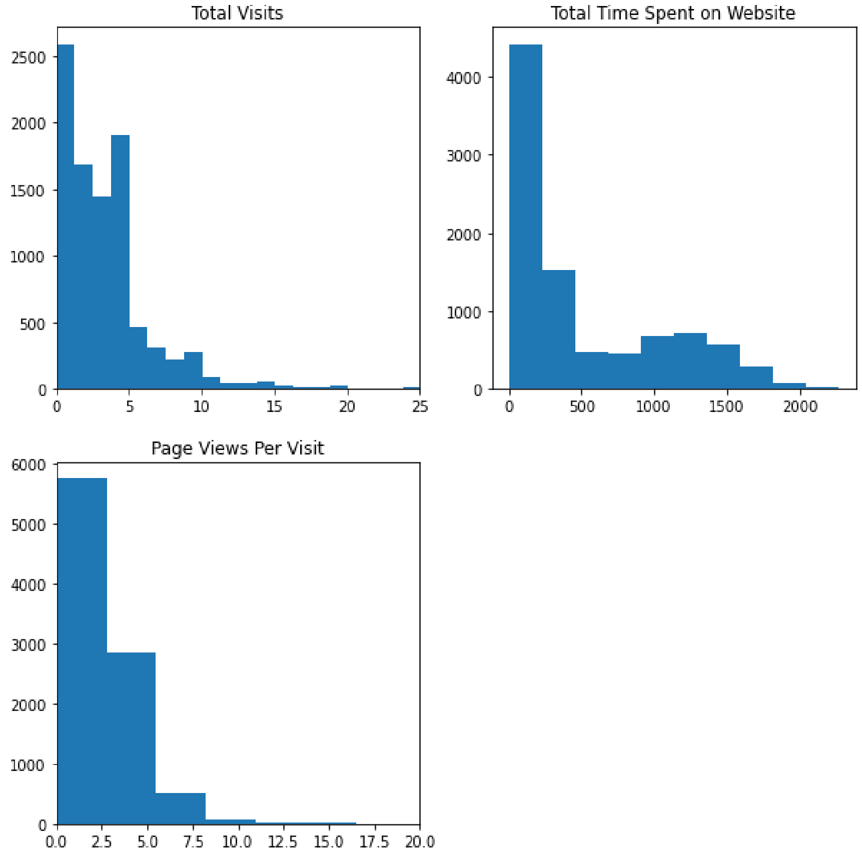
a. Google
b. Direct traffic
c. Organic search etc ..many are converting successful leads

Numerical Analysis

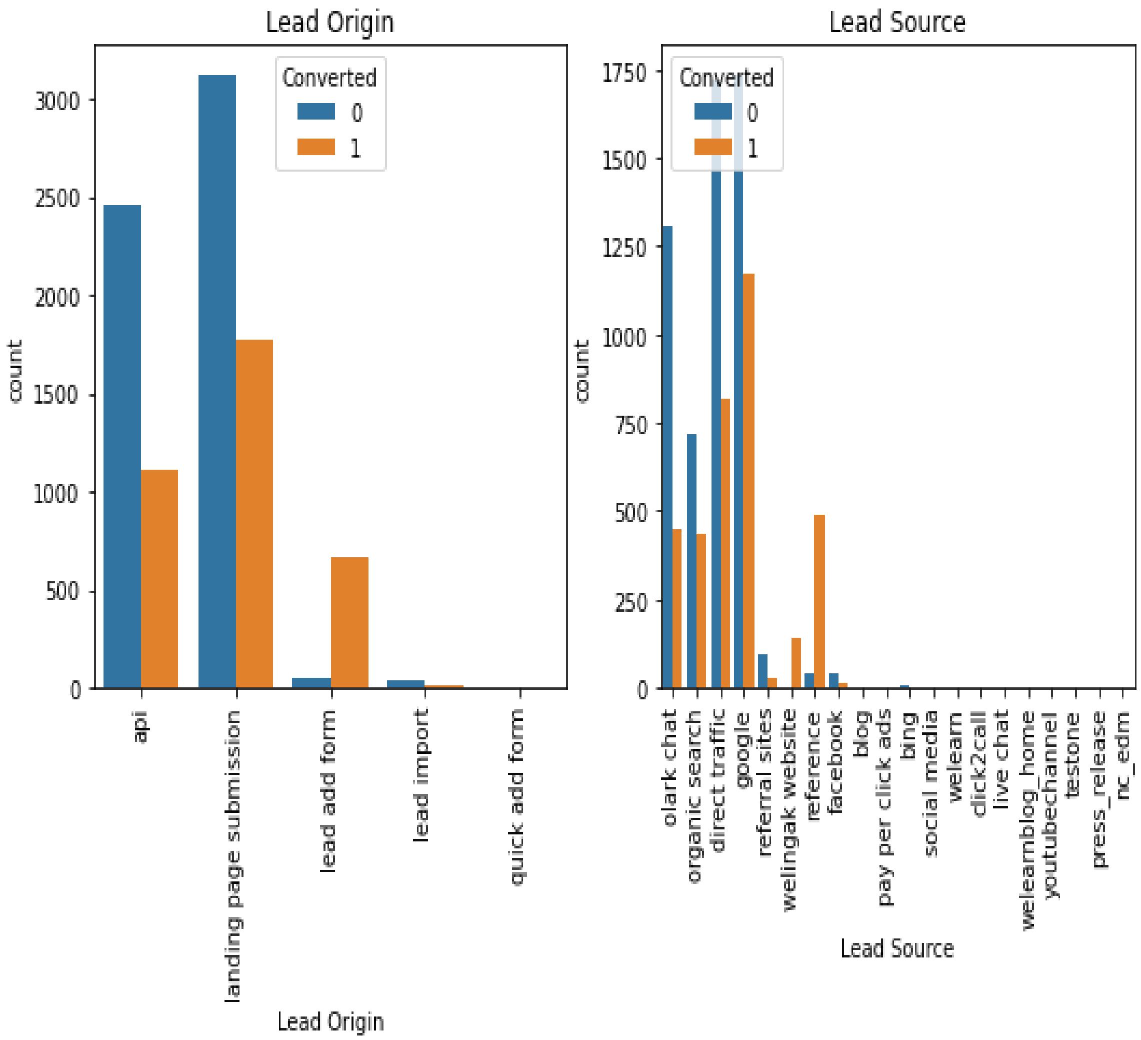
- From Numerical analysis plots we can clearly see

  “Total visits & Total time spent on websites “

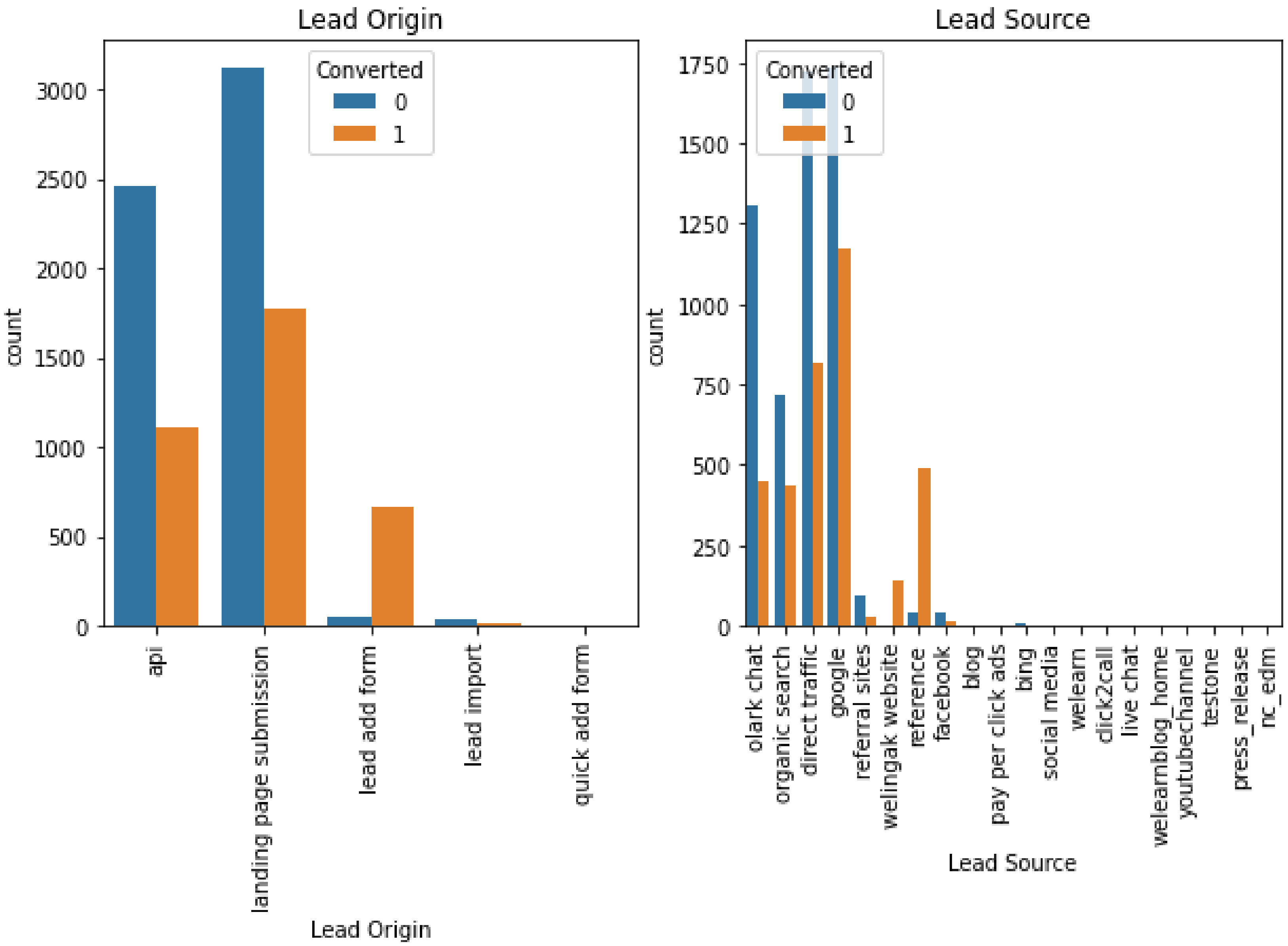
  are top are successful converters



Relating all the categorical variables to Converted



Relating all the categorical variables to Converted

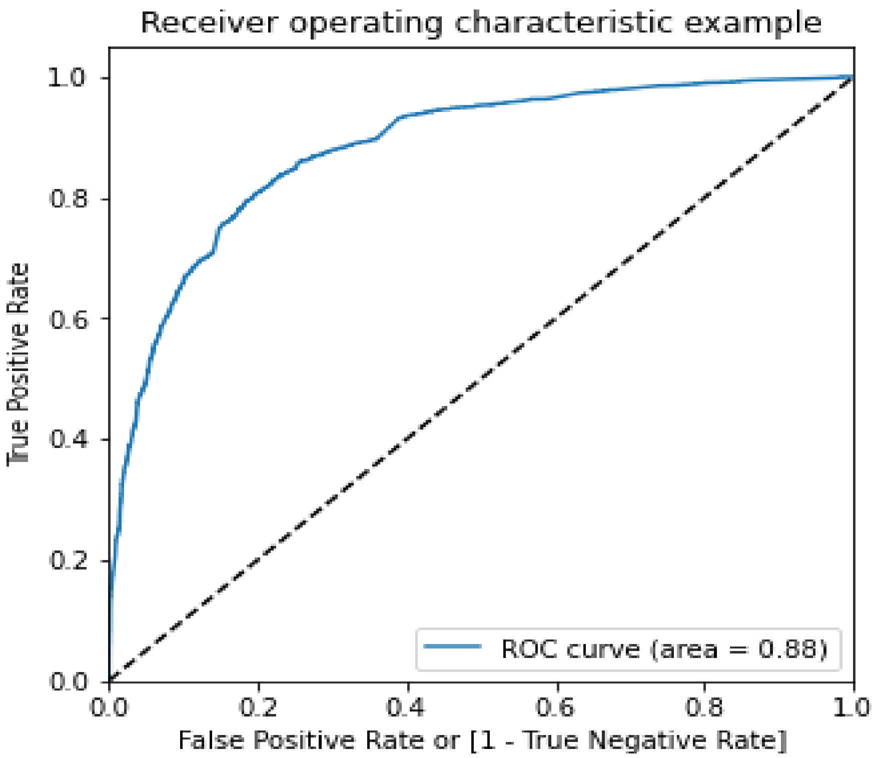


Model Building:

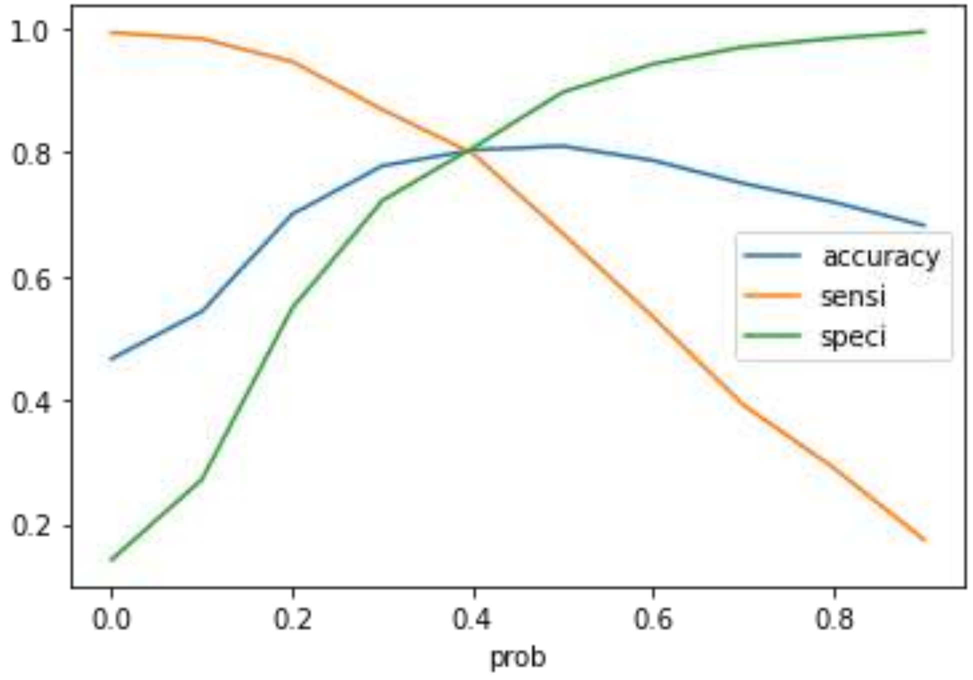
□ Firstly, RFE was done to attain the top 25 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept). Predicting the probabilities on the train set

Model Evaluation:

□ A confusion matrix was made. Check the overall accuracy found 81% Later on the optimum cut off value 0.5 (using ROC curve) was used to find the accuracy, The area under ROC curve is 0.88 which is a very good value.



Optimal Probability Threshold



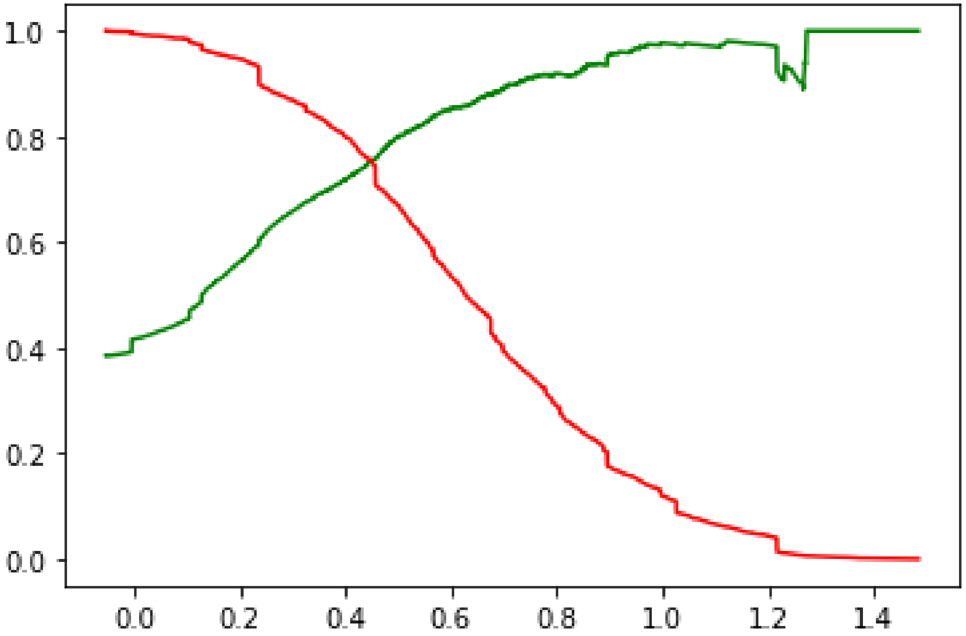
From the graph it is visible that the optimal cut off is at 0.40

Calculating the sensitivity
TP/(TP+FN)=83.62%

Calculating the specificity
TN/(TN+FP) =77.04%

overall accuracy
=79.54%

Precision and recall tradeoff



With the current cut off as 0.42

Precision = TP / TP + FP
TP / (TP + FP)=73.26%

Recall = TP / TP + FN
TP / (TP + FN)=78.02%

Conclusion & Recommendation

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

1. The total time spend on the Website.
2. What is your current occupation housewife.
3. When the lead source was:
a. Google
b. Direct traffic
c. Organic search
4. Lead Origin lead add form.
5. What is your current occupation working professional.
6. Last Notable Activity unreachable
a. modified
b. email opened
c. SMS sent
7. Last Activity others
8. Total Visits
9. What is your current occupation student
10. What is your current occupation unemployed.
11. Last Activity
a. Email opened
b.sms sent
c. olark chat conversation.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

X Education shouldn't focus on the following:

1. Page Views Per Visit
2. Lead Source reference.

Thank you