# Lead Score Assignment

Chitra Rajendran & Sanghati Chatterjee

15.08.2023

# Problem Statement

**Business Understanding**

- 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.

**Business Objective**

- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.

**Data Understanding**

- 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.

# Assumptions:

Basic assumptions that must be met for logistic regression include independence of errors, linearity in the logit for continuous variables, absence of multicollinearity, and lack of strongly influential outliers.

# BUSINESS OBJECTIVE

- The company requires a model to be built for selecting most promising leads.
- Lead score to be given to each leads such that it indicates how promising the lead could be.
- Deployment of the model for future use.

# Solution Steps

1. Data Cleaning and Data Manipulation
- Check and Handle duplicate data
- Check and handle NA and missing values
- Drop columns, if it contains large amount (more than 40%) of missing values and unique values which are not useful for analysis.
- Imputations of the values using Mode.
- Check and handle outliers in Data.

2. EDA
- Univariate Data analysis: value count, distribution of variable etc.
- Bivariate data analysis: Correlation coefficients and pattern between the variables.

3. Feature Scale and Dummy variables creation

4.Classification Techniques: Logistic regression used for the model making and predictions

5.Validation of Model

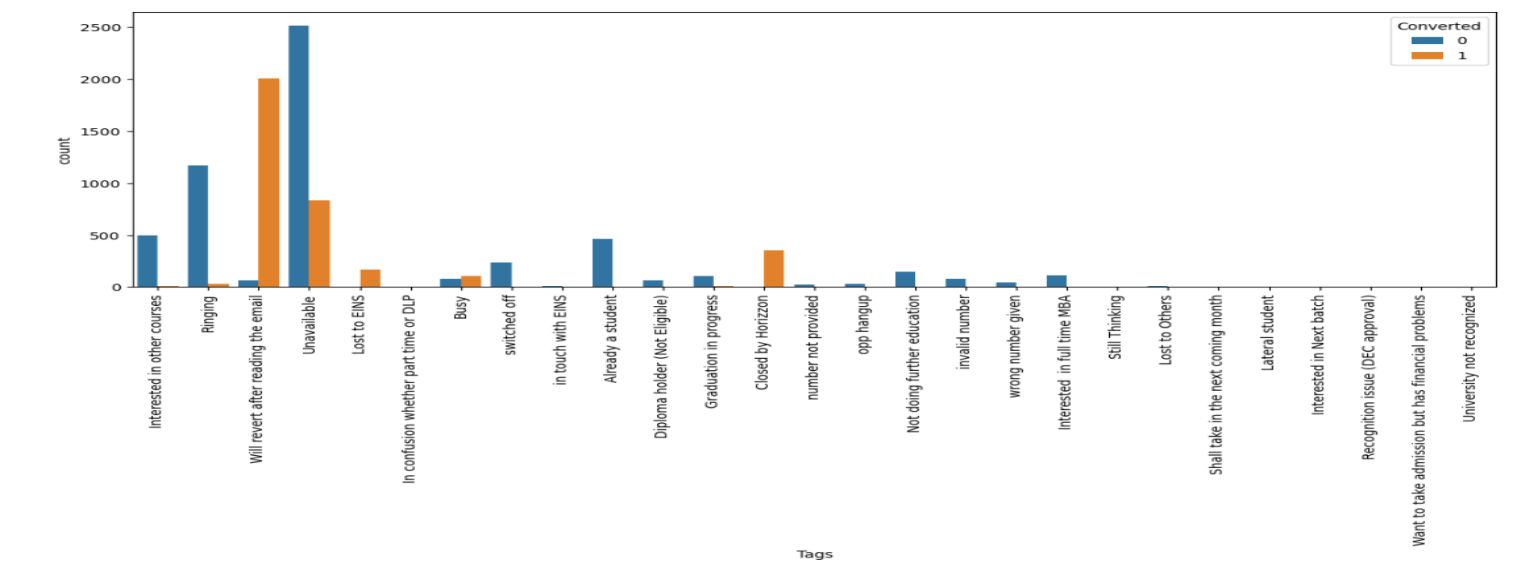
6. Model Presentation

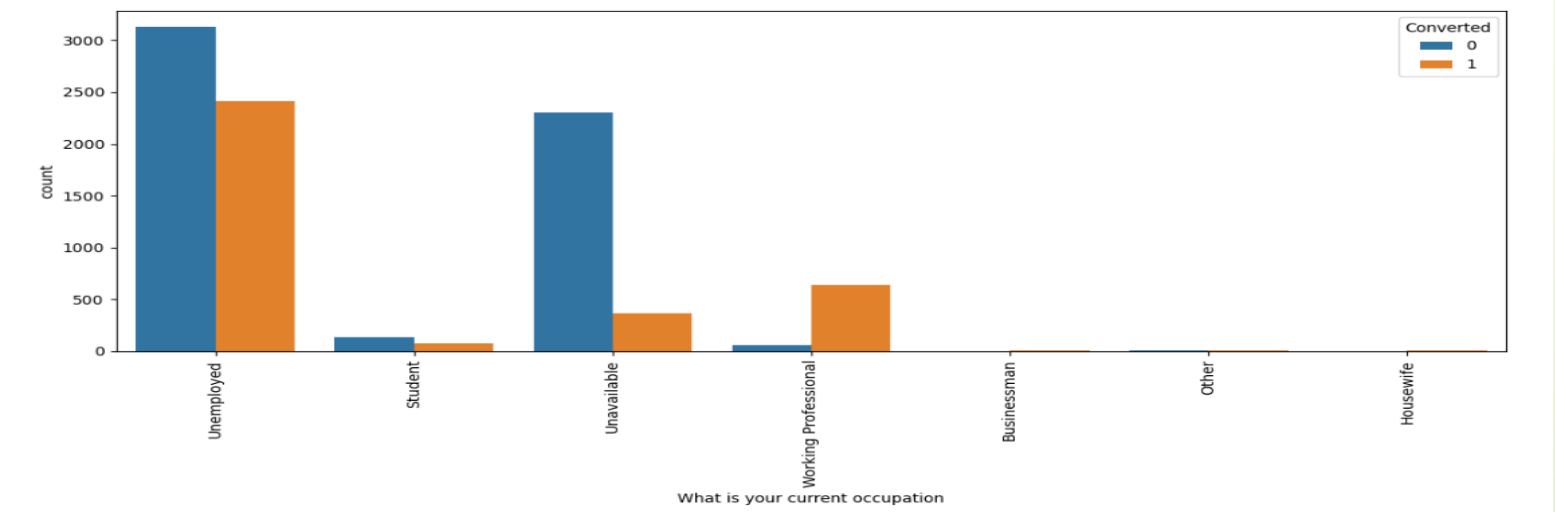
7. Conclusions and recommendations.

# Data Manipulation

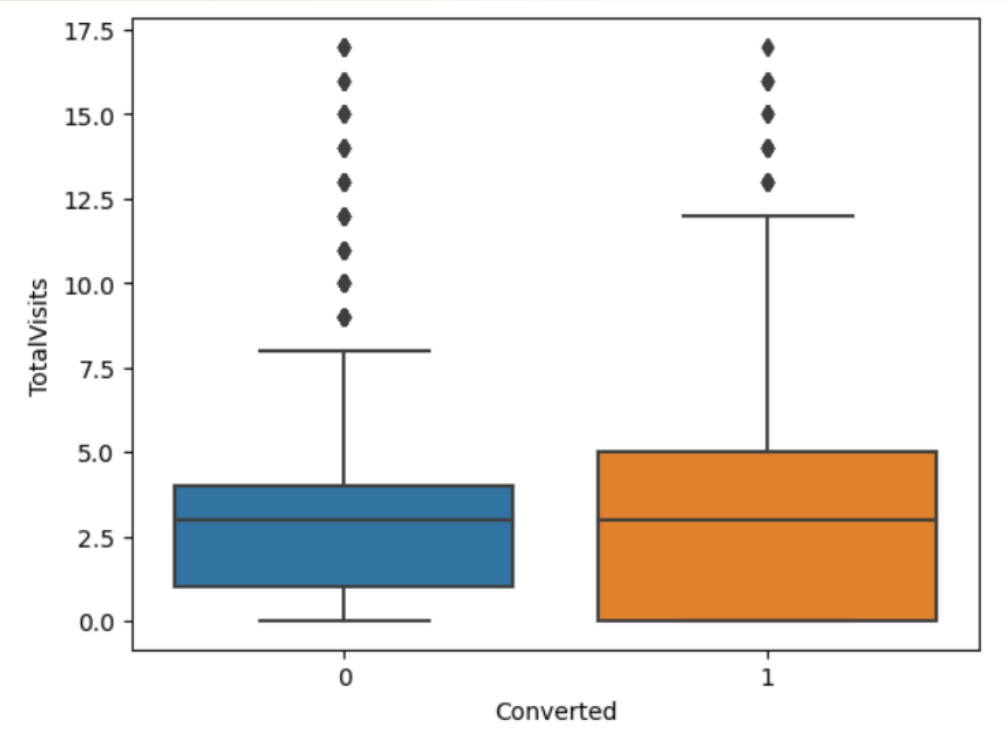
- Total Number of Rows =9240, Total Number of Columns =37.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call", "What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 40% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

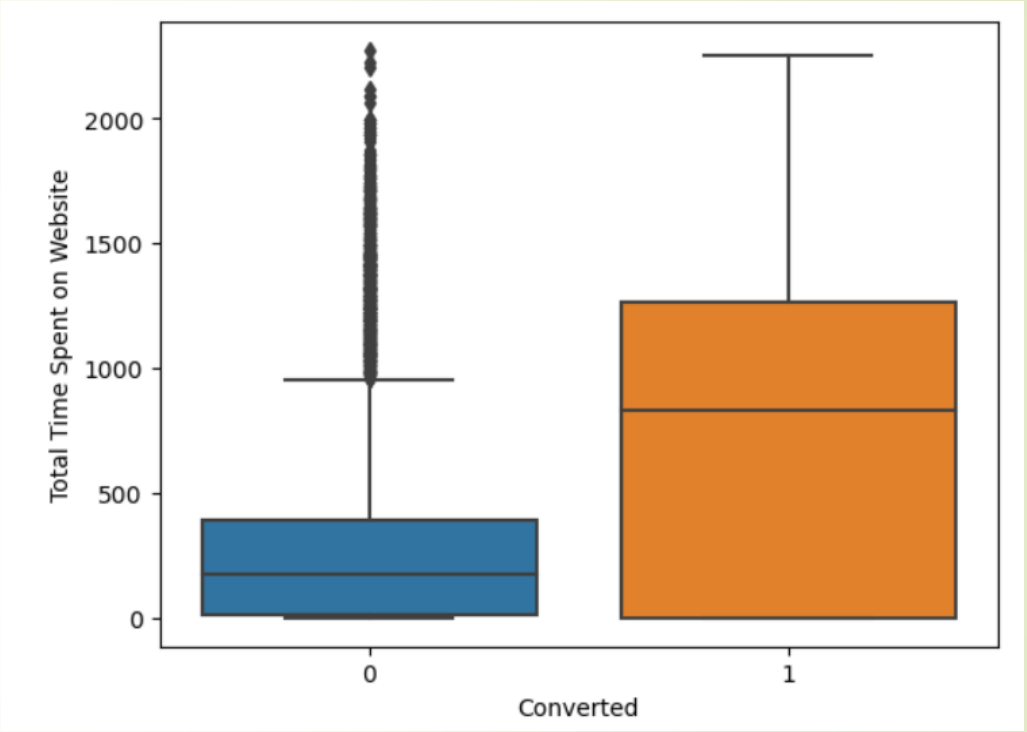
# EDA


Converted
0
1
count
2500
2000
1500
1000
500
0
Interested in other courses
Ringing
Will revert after reading the email
Unavailable
Lost to EINS
In confusion whether part time or DLP
Busy
switched off
in touch with EINS
Already a student
Diploma holder (Not Eligible)
Graduation in progress
Closed by Horizzon
number not provided
opp hangup
Not doing further education
invalid number
wrong number given
Interested in full time MBA
Still Thinking
Lost to Others
Shall take in the next coming month
Lateral student
Interested in Next batch
Recognition issue (DEC approval)
Want to take admission but has financial problems
University not recognized
Tags

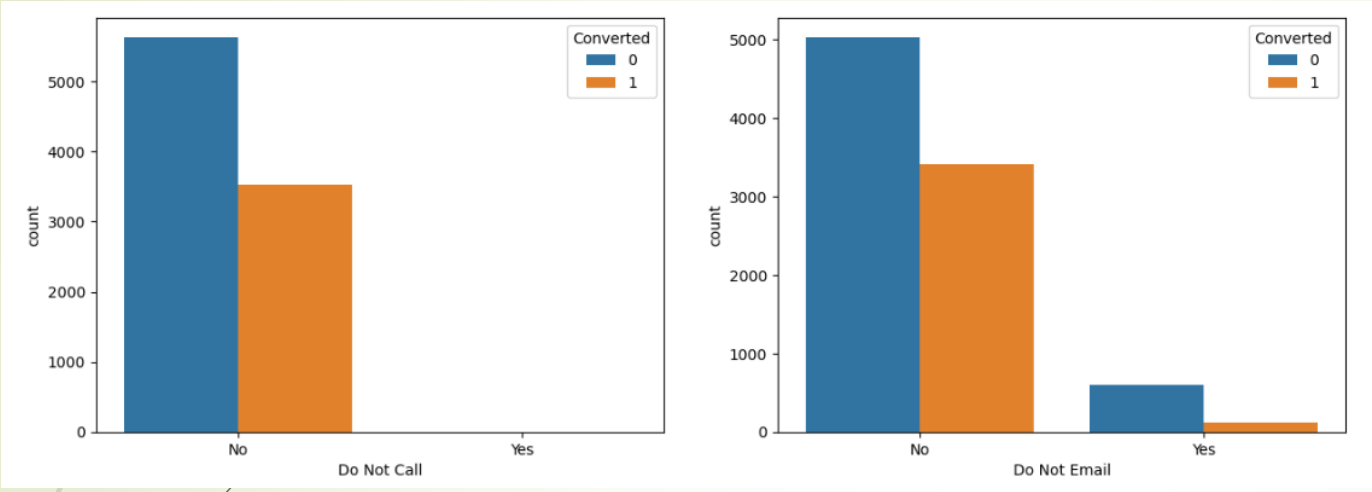

Converted
0
1
count
3000
2500
2000
1500
1000
500
0
Unemployed
Student
Unavailable
Working Professional
Businessman
Other
Housewife
What is your current occupation

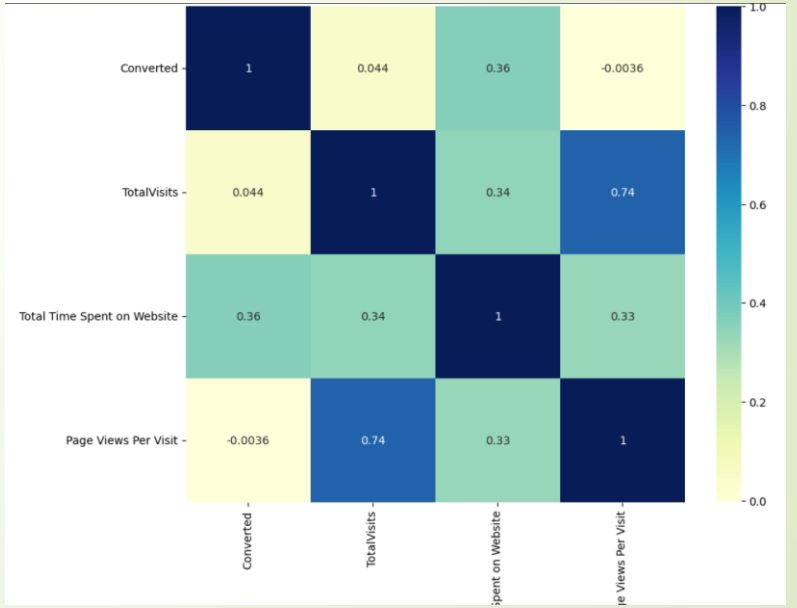
# EDA


17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
TotalVisits
0
1
Converted


2000
1500
1000
500
0
Total Time Spent on Website
0
1
Converted

# EDA


Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Call
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email


Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1
0.044
0.36
-0.0036
0.044
1
0.34
0.74
0.36
0.34
1
0.33
-0.0036
0.74
0.33
1
1.0
0.8
0.6
0.4
0.2
0.0
Converted
TotalVisits

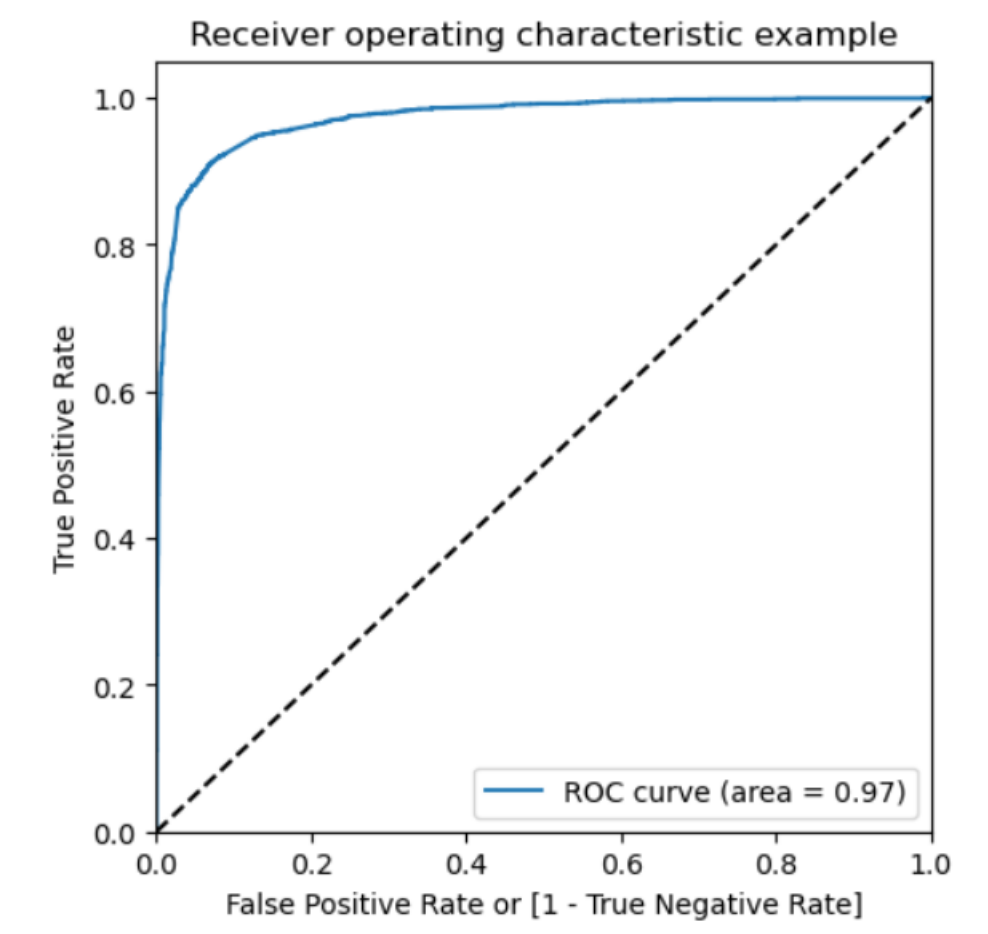
# Data Conversion

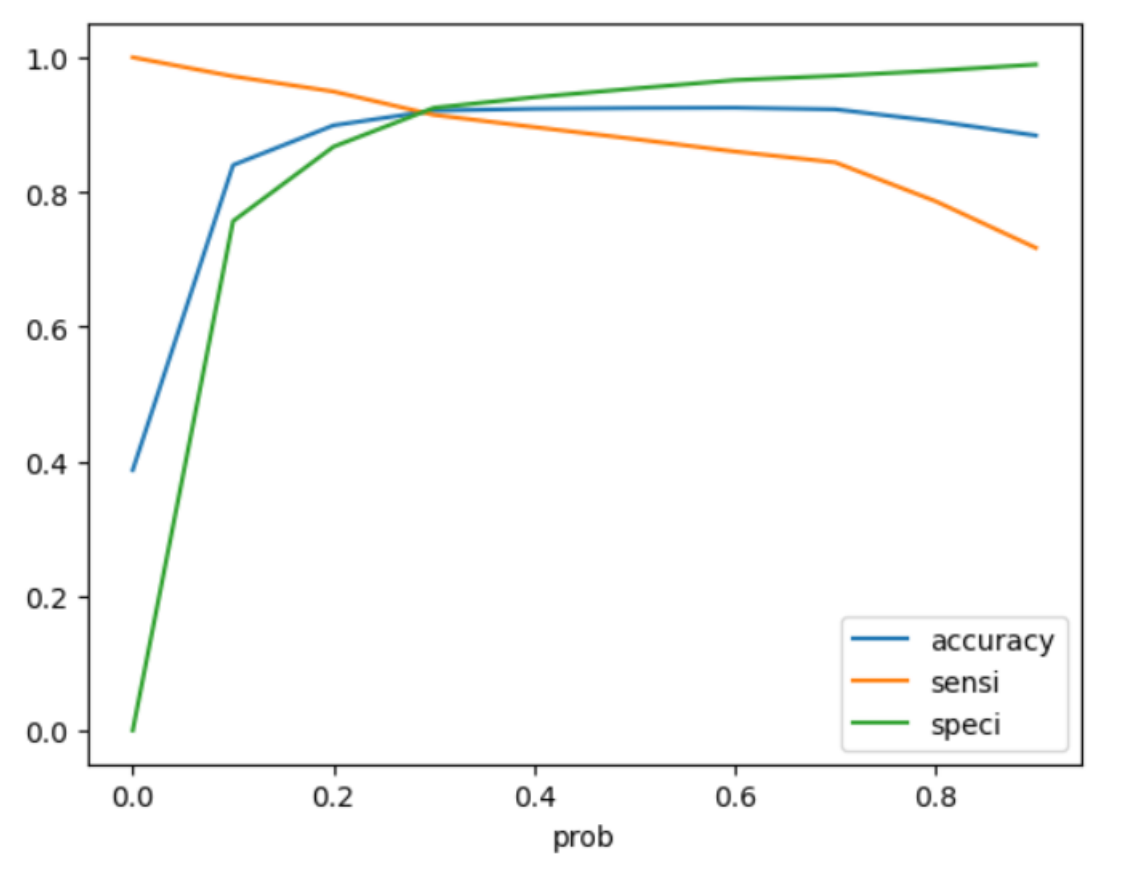
- Numerical Variables are Normalised.
- Dummy Variables are created for object type variables.
- Total Rows for Analysis: 9157
- Total Columns for Analysis: 78

# Model Building

- Splitting the Data into Training and Testing Sets.
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection.
- Running RFE with 76 variables as output (with the variables created after creating dummy variables)
- Building Model by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 3.
- Predictions on test data set.
- Overall accuracy 93.3%.
- Sensitivity is 89.7%.
- Specificity is 95%.

# ROC Curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.97)
0.0
0.2
0.4
0.6
0.8
1.0


accuracy
sensi
speci
prob
0.0
0.2
0.4
0.6
0.8
1.0

- Finding Optimal Cut off Point

- Optimal cut off probability is that probability where we get balanced sensitivity and specificity.

- From the second graph it is visible that the optimal cut off is approximately at 0.34.

**Top Columns based on the Final columns selected in the Logistic Regression are:**

- Tags_Will revert after reading the email
- Total Time Spent on Website
- Last Notable Activity_SMS Sent

# Conclusion

It was found that the variables that mattered the most in the potential buyers are:

- Tags_Will revert after reading the email
- The Total Time Spent on Website
- When the Last Activity was :
  - SMS
  - OlarkChat Conversation
- Total number of Visits
- When The Lead source was
  - Olark Chat
  - Wellingak Website
- When the lead origin is Lead add Form.
- When the current occupation was :
  - Working Professionals
  - Student
  - Unemployed
  - Other

Keeping the above mentioned points in mind the X education can increase all the potential buyers to change their mind and buy their courses.