LEAD SCORING

Case Study

By Pawan & Suraaj

CONTENTS

PROBLEM STATEMENT

- X Education sells online courses to industry professionals. They get lead through referrals, website forms and other sources. The typical lead conversion rate at X education is around 30%. 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.

OBJECTIVE

- 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.

APPROACH ANALYSIS

- Knowing the data
- Cleaning the data
- Outlier treatment
- Exploratory data analysis
- Building & Evaluating a Logistic Regression model
- Assigning a score to classify a ‘Hot Lead’

INSIGHTS ON DATA

- Missing Data
  - Actions taken: Dropping the row

Columns	Missing Value %
Lead Source	0.39
TotalVisits	1.48
Page Views Per Visit	1.48
Last Activity	1.11

  - Actions taken: Imputation

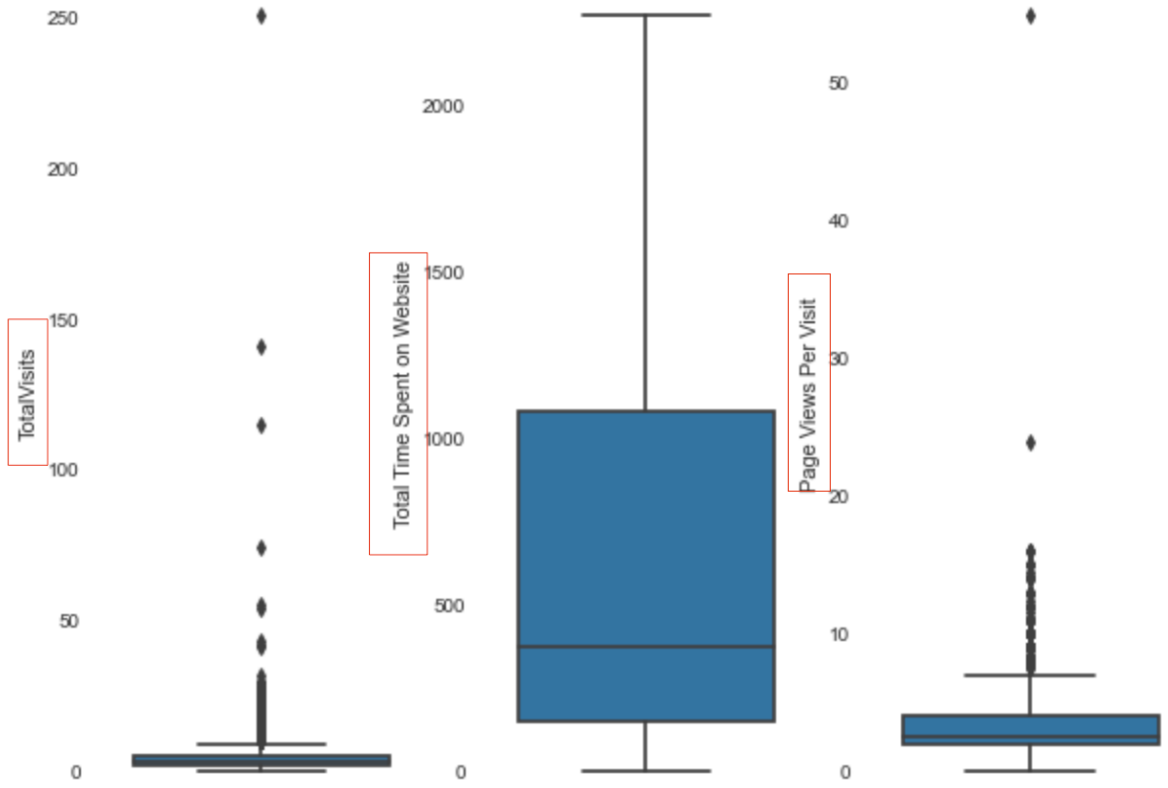
Columns	Missing Value %
Specialization	15.562771
City	15.367965

- Highly skewed data (Action taken: Dropping columns)
  - 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', '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', 'Do Not Call', 'Do Not Email', 'Country'

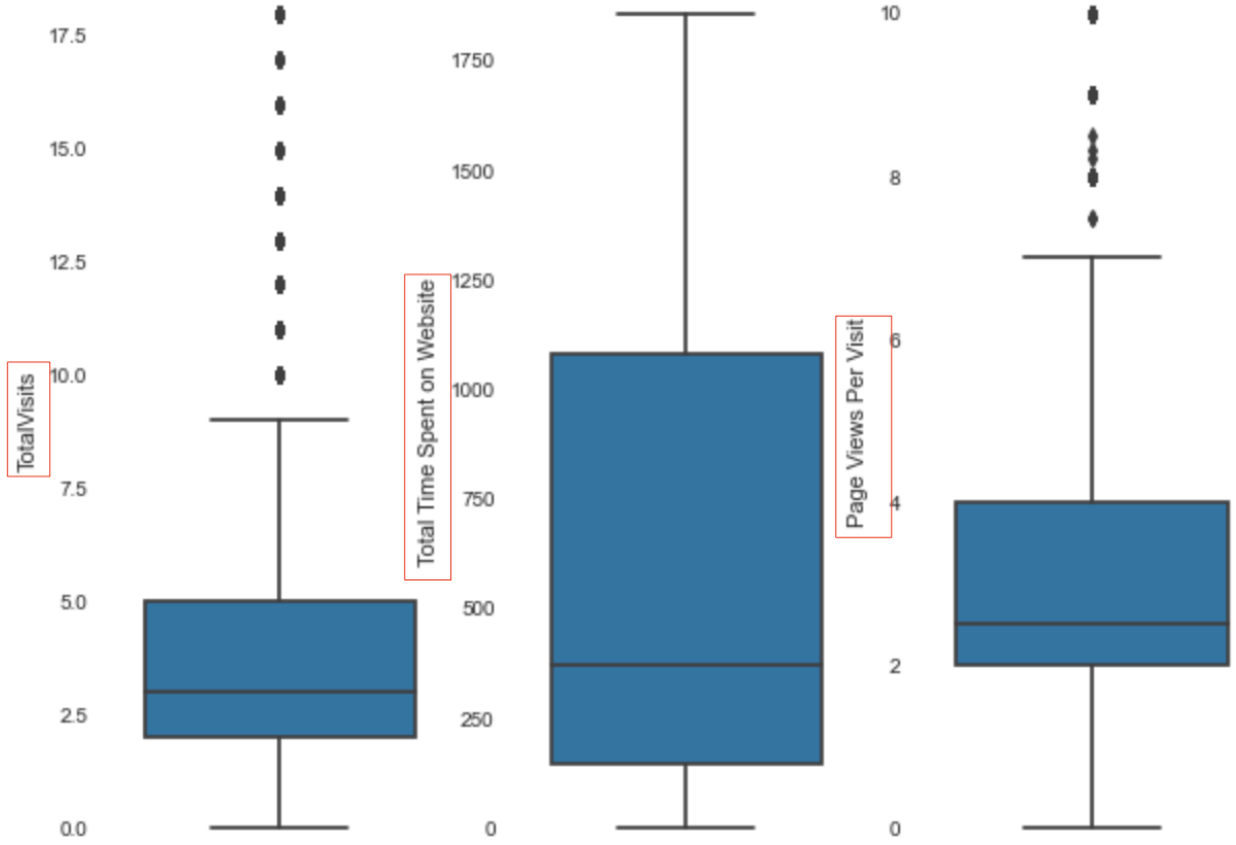
OUTLIER TREATMENT

CHECKING FOR OUTLIERS IN THE CONTINUOUS VARIABLES (KEEPING THE DATA WITHIN IQR)

BEFORE TREATMENT



AFTER TREATMENT



MULTICOLLINEARITY

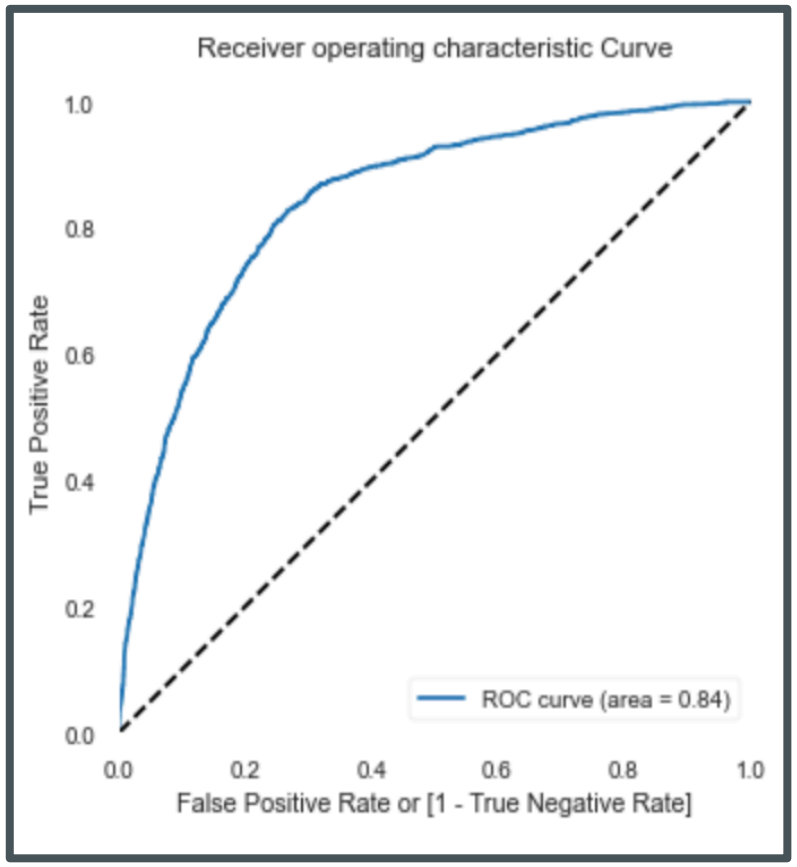
Note: We are not dropping high correlation values here. Will do it in later steps using RFE & VIF

level_0	level_1	Correlation
Last Notable Activity_Email Marked Spam	Last Activity_Email Marked Spam	1.000000
Lead Source_Facebook	Lead Origin_Lead Import	0.970046
Lead Source_Reference	Lead Origin_Lead Add Form	0.890326
Last Notable Activity_Email Opened	Last Activity_Email Opened	0.866782
Last Notable Activity_Unsubscribed	Last Activity_Unsubscribed	0.861715
Last Notable Activity_SMS Sent	Last Activity_SMS Sent	0.850708
ast Notable Activity_Had a Phone Conversation	Last Activity_Had a Phone Conversation	0.844818
Last Notable Activity_Email Link Clicked	Last Activity_Email Link Clicked	0.784597
Last Notable Activity_Page Visited on Website	Last Activity_Page Visited on Website	0.708499
Last Notable Activity_Email Received	Last Activity_Email Received	0.707037

DATA PREPARATION & MODEL BUILDING

- Rescaling the training variables
- Feature Selection Using RFE
- Assessing the model with StatsModels
- Dropping High P- Value and High VIF variables
- ROC & Precision- Recall Curves
- Finding optimal cut-off
- Evaluating the model using metrices like Accuracy Score, Sensitivity, Specificity, Precision

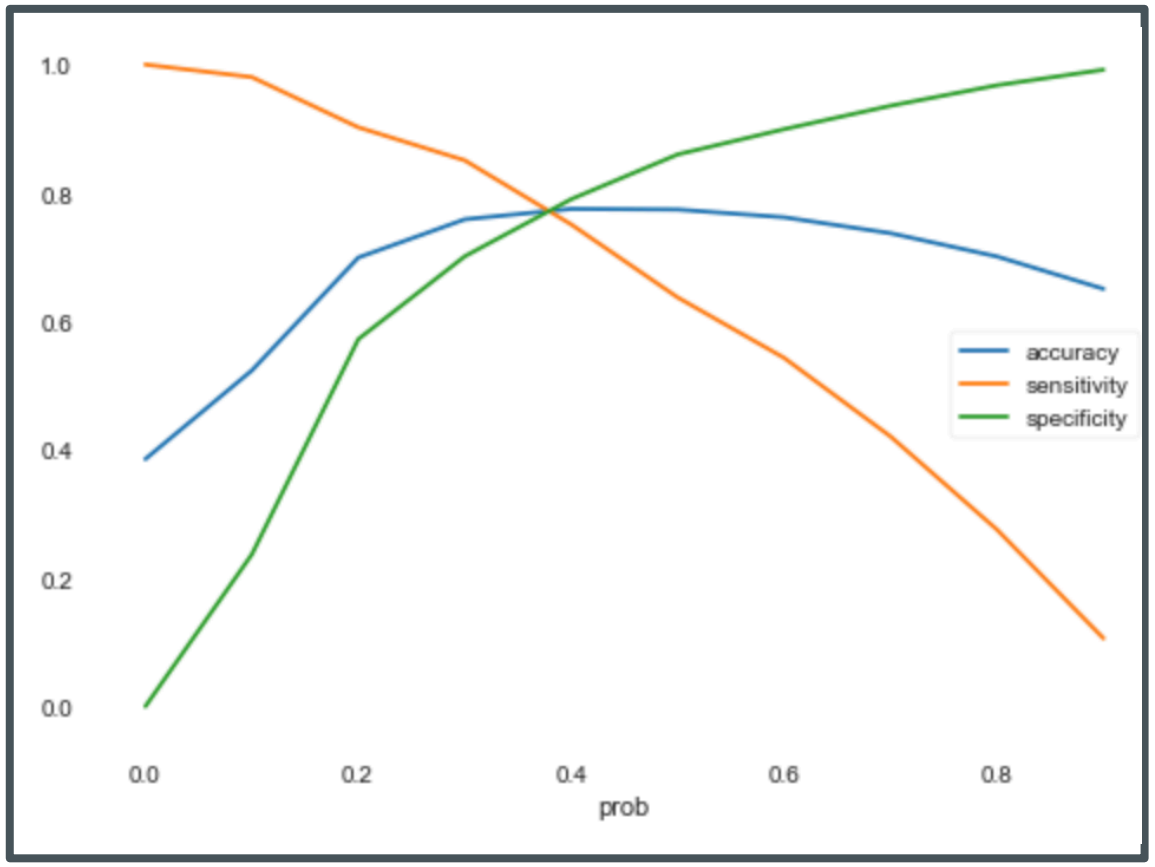
PLOTTING THE ROC CURVE



An ROC curve demonstrates several things:

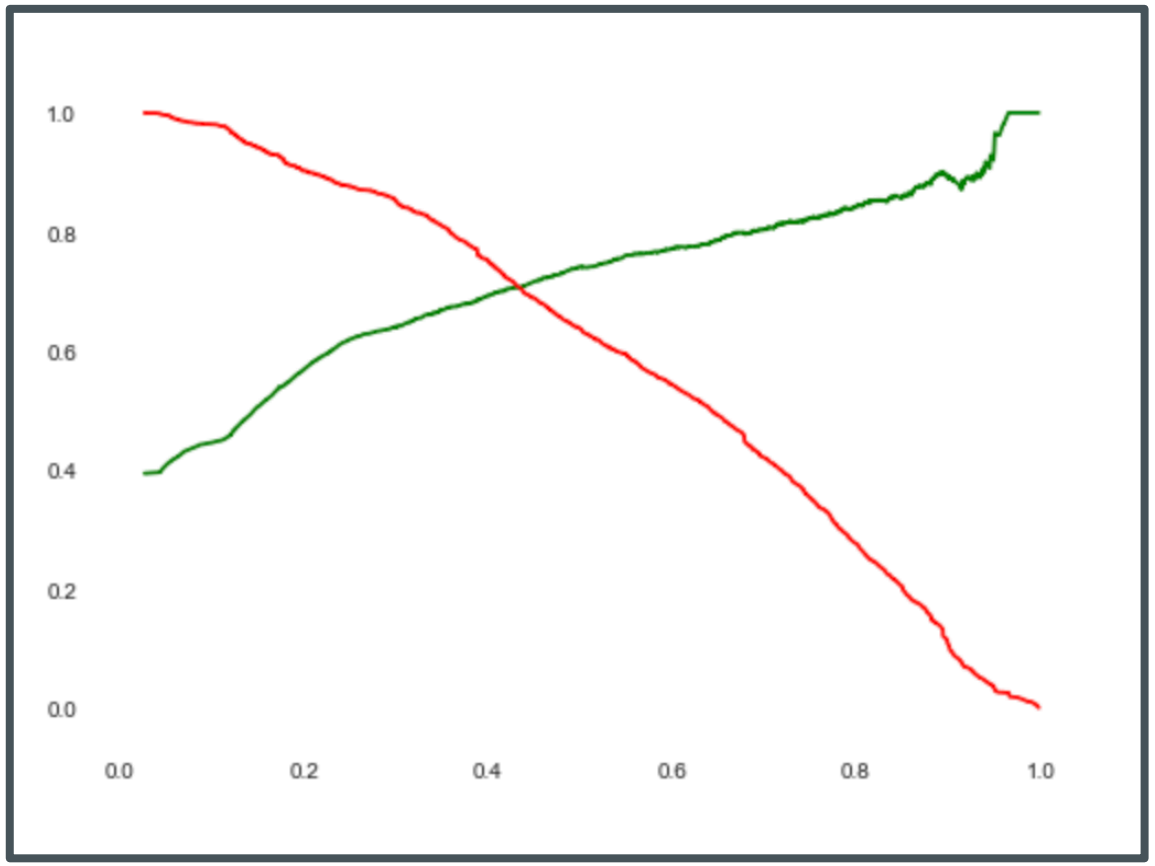
1. It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
2. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
3. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

FINDING THE OPTIMAL CUT-OFF



- Plotting accuracy sensitivity and specificity for various probabilities
- **From the curve, 0.37 is the optimum point to take it as a cutoff probability.**

PRECISION AND RECALL TRADEOFF



Green- Precision Red- Recall

Whether to Use ROC or Precision-Recall Curves?

- Generally, the use of ROC curves and precision-recall curves are as follows:

1. ROC curves should be used when there are roughly equal numbers of observations for each class.
2. Precision-Recall curves should be used when there is a moderate to large class imbalance.

PREDICTION ON TEST SET & EVALUATION

- Model gave promising results:

1. Accuracy: 0.78
2. Sensitivity: 0.79
3. Specificity: 0.77
4. Precision: 0.67
5. Actual Conversion Ratio: 38.5%
6. Model Predicted Conversion Ratio: 39%