Lead Scoring Case Study

Problem Statement

An education company named X Education sells online courses to industry professionals. Leads are potentially generated through multiple sources as provided in the dataset. The typical lead conversion rate at X education is around 30%. The company is looking to make its process more efficient, whereby it can identify and focus on the most potential leads, with a target to achieve the Lead Conversion Rate to 80%

Lead Conversion Process

Desired Solution

Develop a Logistic Regression Model to assign a lead score between 0 – 100 for each of the leads which can be used by the Company to target potential leads. Identify the key factors that drive convertibility

Data Provided

Variables	Description
Prospect ID	A unique ID with which the customer is identified.
Lead Number	A lead number assigned to each lead procured.
Lead Origin	The origin identifier with which the customer was identified to be a lead. Includes API, Landing Page Submission, etc.
Lead Source	The source of the lead. Includes Google, Organic Search, Olark Chat, etc.
Do Not Email	An indicator variable selected by the customer wherein they select whether of not they want to be emailed about the course or not.
Do Not Call	An indicator variable selected by the customer wherein they select whether of not they want to be called about the course or not.
Converted	The target variable. Indicates whether a lead has been successfully converted or not.
TotalVisits	The total number of visits made by the customer on the website.
Total Time Spent on Website	The total time spent by the customer on the website.
Page Views Per Visit	Average number of pages on the website viewed during the visits.
Last Activity	Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.
Country	The country of the customer.
Specialization	The industry domain in which the customer worked before. Includes the level 'Select Specialization' which means the customer had not selected this option while filling the form.
How did you hear about X Education	The source from which the customer heard about X Education.
What is your current occupation	Indicates whether the customer is a student, umemployed or employed.
What matters most to you in choosing this course	An option selected by the customer indicating what is their main motto behind doing this course.
Search	Indicating whether the customer had seen the ad in any of the listed items.
Magazine	
Newspaper Article	
X Education Forums	
Newspaper	
Digital Advertisement	
Through Recommendations	Indicates whether the customer came in through recommendations.
Receive More Updates About Our Courses	Indicates whether the customer chose to receive more updates about the courses.
Tags	Tags assigned to customers indicating the current status of the lead.
Lead Quality	Indicates the quality of lead based on the data and intuition the the employee who has been assigned to the lead.
Update me on Supply Chain Content	Indicates whether the customer wants updates on the Supply Chain Content.
Get updates on DM Content	Indicates whether the customer wants updates on the DM Content.
Lead Profile	A lead level assigned to each customer based on their profile.
City	The city of the customer.
Asymmetrique Activity Index	An index and score assigned to each customer based on their activity and their profile
Asymmetrique Profile Index	
Asymmetrique Activity Score	
Asymmetrique Profile Score	
I agree to pay the amount through cheque	Indicates whether the customer has agreed to pay the amount through cheque or not.
a free copy of Mastering The Interview	Indicates whether the customer wants a free copy of 'Mastering the Interview' or not.
Last Notable Activity	The last notable acitivity performed by the student.

Analysis Approach - 1

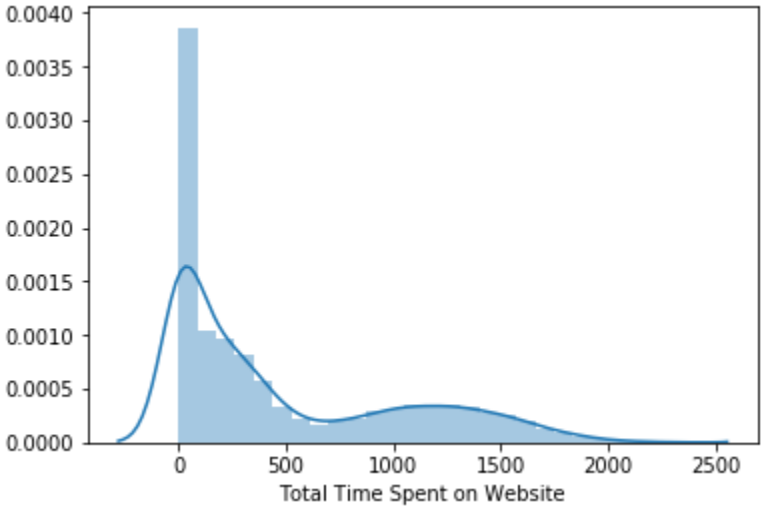
Perform EDA on the DataSet

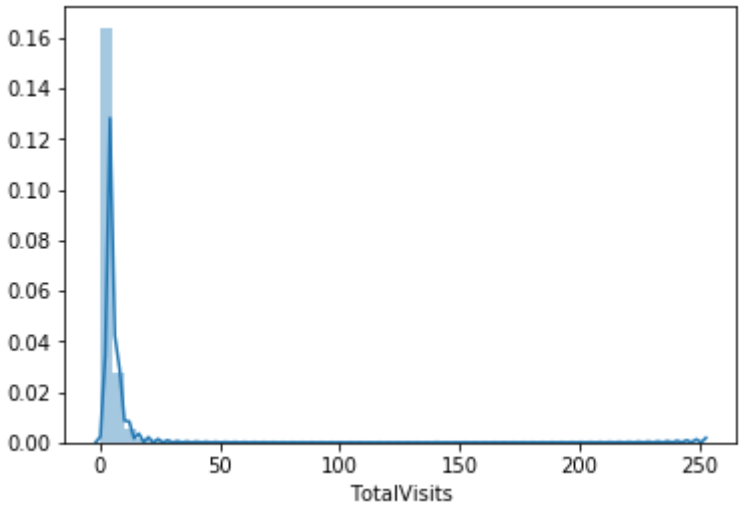
- Handling Null values and ‘Select’ Values
- Handling values which had a very low count – Combining such smaller values into a single value
- Finding Relationship between columns
- Dropping the columns with a high percentage of missing values
- Dropping the columns which had either a single value OR binary values with <10 occurrences of the other value
- Converting the Remaining binary columns (From Yes/No to 1/0)

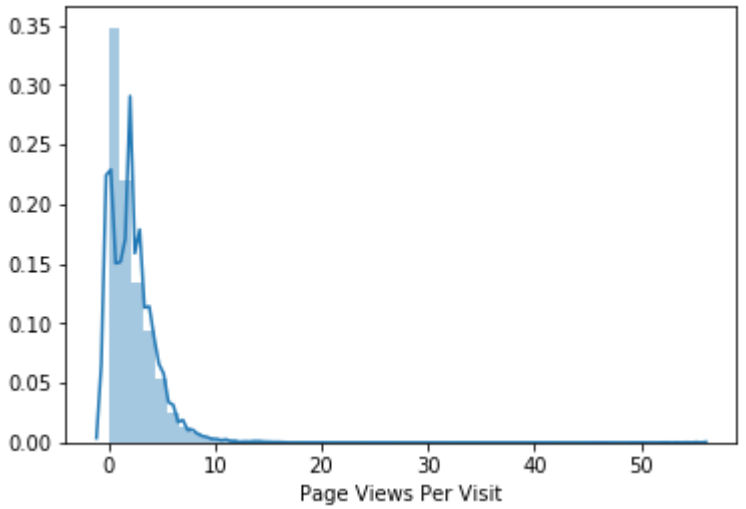
Result :
# of Rows Remaining (9103) [# of Rows in Initial Dataset = 9240]
# of Columns Remaining (14) [# of Columns in Initial Dataset = 37]

Analysis Approach - 2

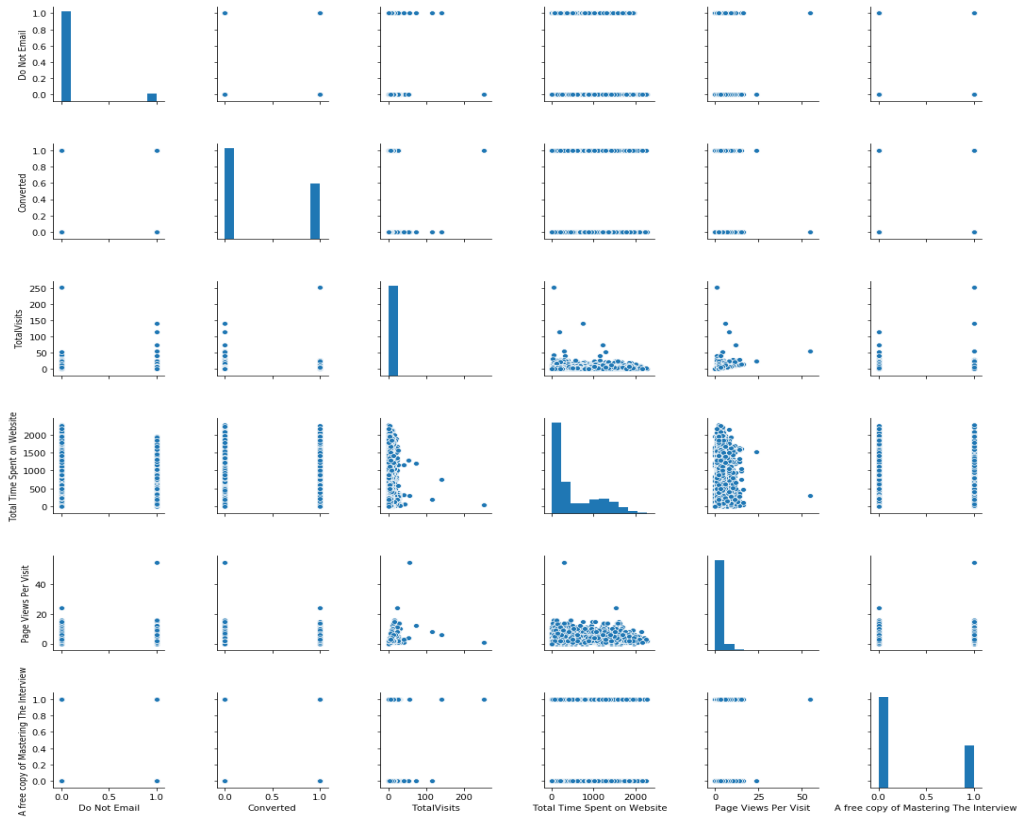
Visualize Numerical Attribute Distribution using DistPlot

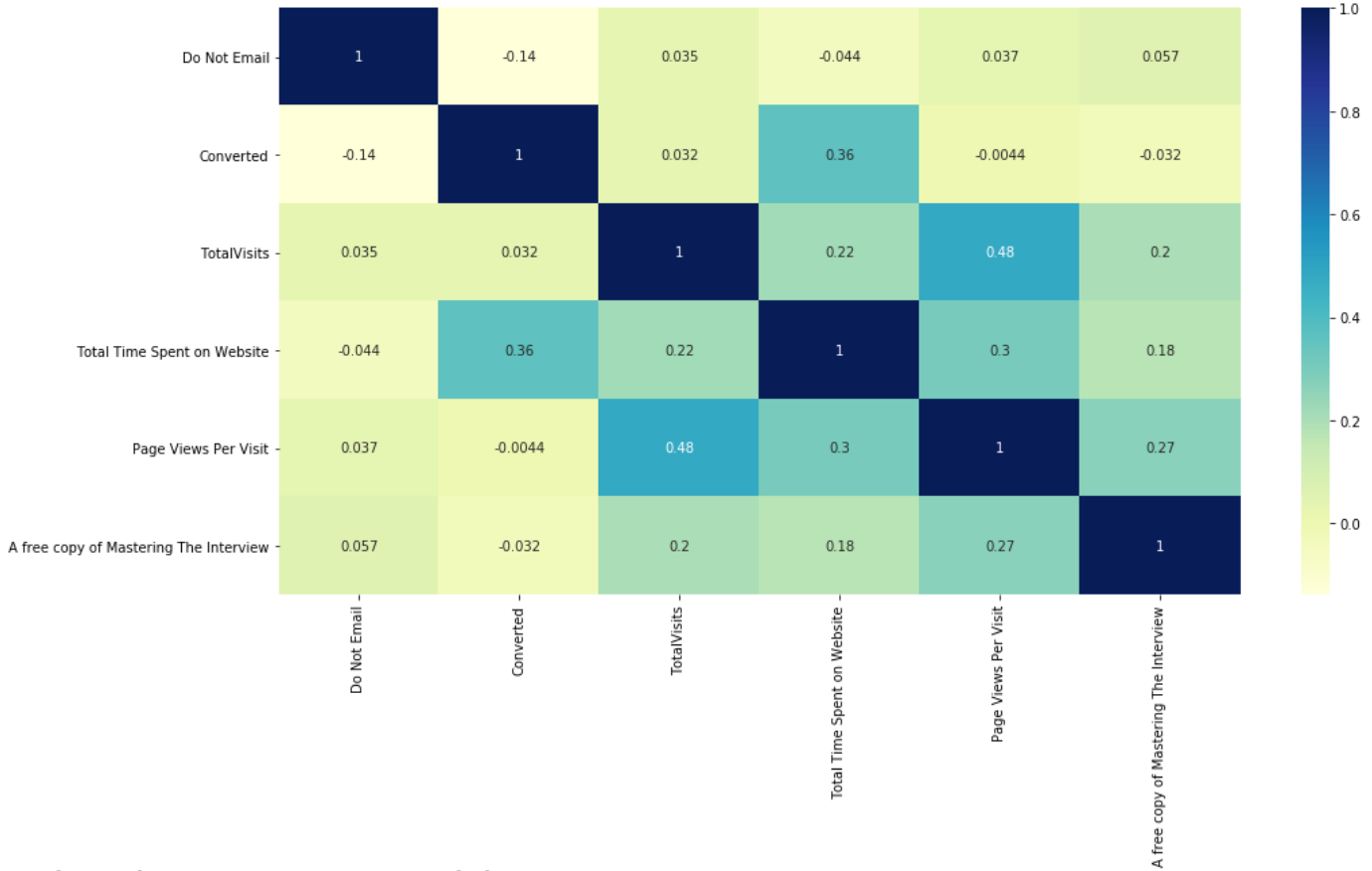






Analysis Approach - 3





Study Relationship between Variables

Findings :

- `converted` is highly (positively) correlated with `Total Time spent on website`

- `converted` is most strongly negatively correlated to `Do not Email`. This suggest that those prospects who do not want to be contacted by Email have lower probability of conversion

Analysis Approach - 4

Dealing with Categorical Variables

- Convert Categorical variables into Dummy variables

Model Building

- Split the Data into Train set (70%) and Test Set (30%)
- Feature Scaling for Numeric Variables
- Feature Selection using RFE
- Assessing the model with StatsModels
- Getting the predicted values on the train set
- Creating a dataframe with the actual Convertibility flag and the predicted probabilities
- Checking VIFs (Discarding variables with high VIFs)
- Calculate Accuracy, Specificity, Sensitivity
- Plotting the ROC Curve
- Plotting the Accuracy, Sensitivity, Specificity curve to arrive at the optimum cut-off probability
- Plotting the Precision and Recall curve

Analysis Approach - 5

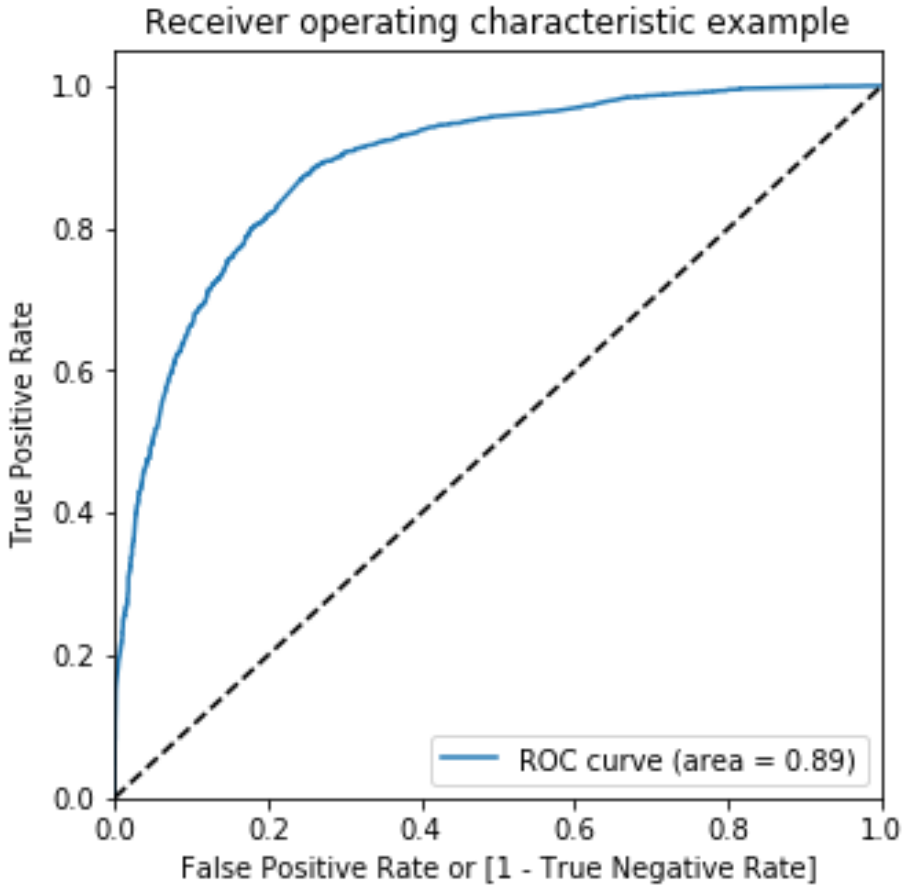
	Features	VIF
2	Lead Origin_Lead Add Form	73.49
4	Lead Source_Reference	57.35
5	Lead Source_Welingak Website	17.20
8	Last Activity_SMS Sent	5.56
14	Last Notable Activity_SMS Sent	5.05
0	Do Not Email	1.86
3	Lead Source_Olark Chat	1.82
6	Last Activity_Email Bounced	1.73
10	What is your current occupation_Unemployed	1.67
7	Last Activity_Olark Chat Conversation	1.47
12	What is your current occupation_unknown	1.46
1	Total Time Spent on Website	1.31
11	What is your current occupation_Working Profes...	1.24
13	Last Notable Activity_Others	1.14
9	What is your current occupation_Student	1.03

VIF Initial

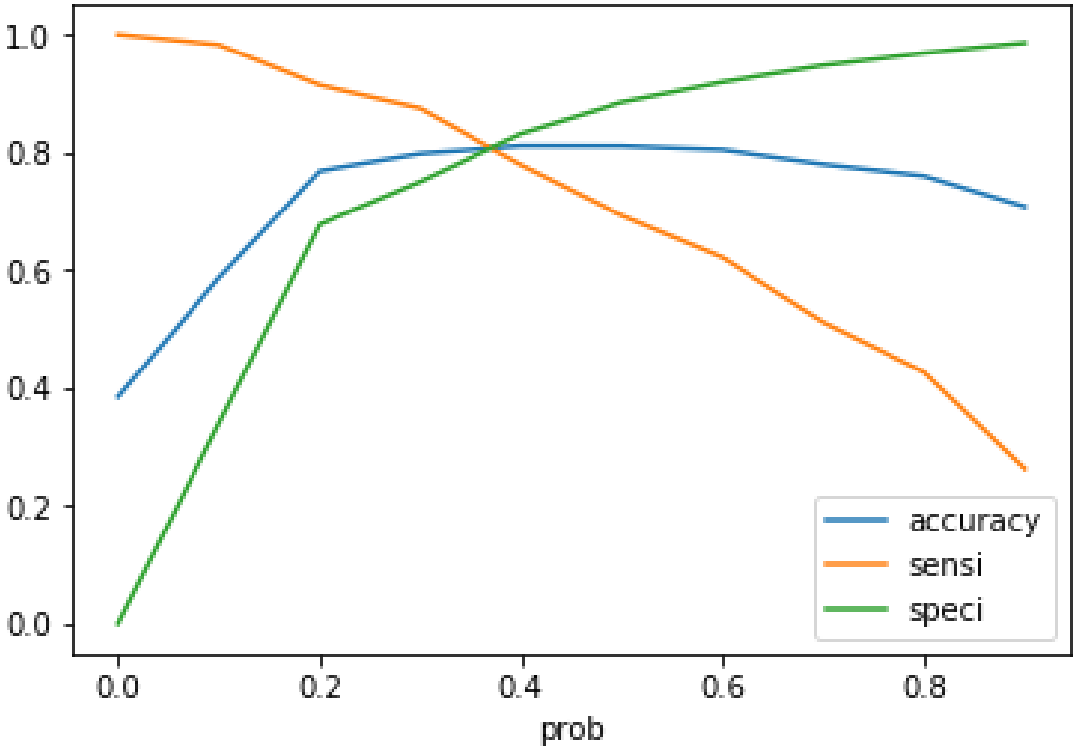
	Features	VIF
7	Last Activity_SMS Sent	5.56
13	Last Notable Activity_SMS Sent	5.05
0	Do Not Email	1.86
2	Lead Source_Olark Chat	1.82
5	Last Activity_Email Bounced	1.73
9	What is your current occupation_Unemployed	1.66
6	Last Activity_Olark Chat Conversation	1.47
11	What is your current occupation_unknown	1.46
1	Total Time Spent on Website	1.31
3	Lead Source_Reference	1.23
10	What is your current occupation_Working Profes...	1.23
12	Last Notable Activity_Others	1.14
4	Lead Source_Welingak Website	1.06
8	What is your current occupation_Student	1.03

VIF Final

Analysis Approach - 6



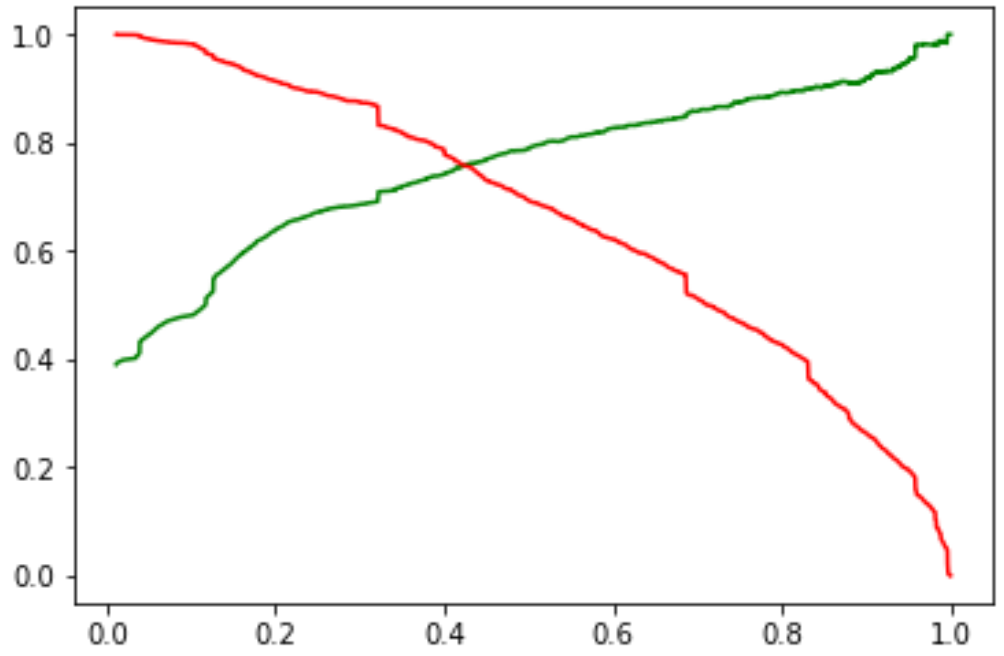
ROC Curve



Accuracy – Sensitivity-Specificity Curve

From the curve, we take 0.4 as the optimum point as the cut-off probability

Analysis Approach - 7



Precision and Recall Curve

Analysis Approach - 8

Assign a Lead Score

- Lead Score = (Converted Probability) * 100 [rounded off to the integer value]

A Lead Score of 56 and above should result in a Hit Rate of 80%

Score Range	
0-10	3.123475
11-20	11.549165
21-30	25.948104
31-40	42.663379
41-50	50.511945
51-60	57.674419
61-70	70.718232
71-80	68.905473
81-90	85.112782
91-100	92.916667

Score Range New	
0-55	18.957871
56-100	80.853352

Convertibility Ratio for Lead Score Range (based on the concept of Binning)