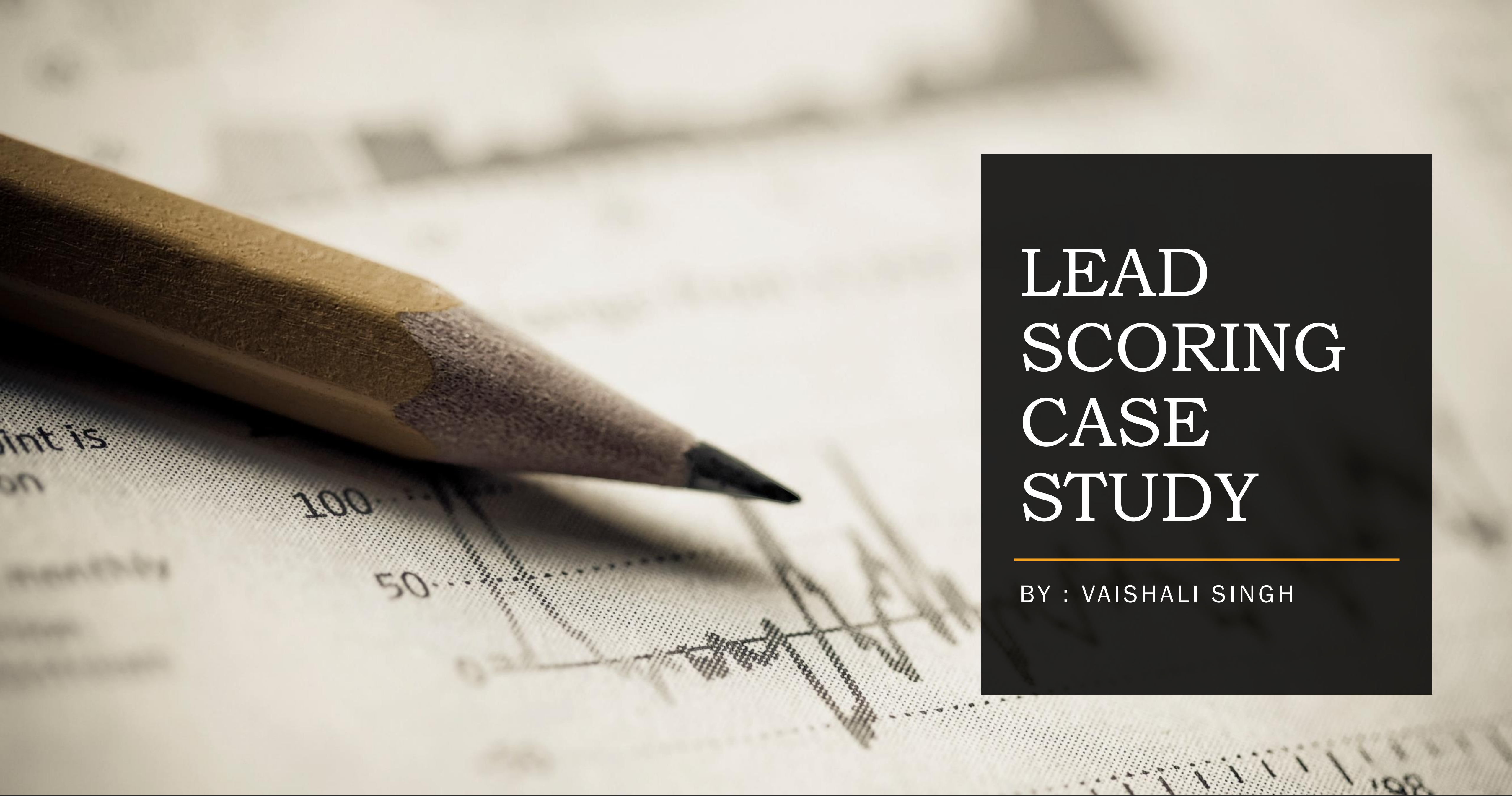


LEAD SCORING CASE STUDY

BY : VAISHALI SINGH

Problem Statement

An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Although X Education gets a lot of leads, its lead conversion rate is very poor. o 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.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Goals of the Case Study

There are quite a few goals for this case study:

1. 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.

2. 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. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Results Expected

1. A well-commented Jupyter notebook with at least the logistic regression model, the conversion predictions and evaluation metrics.

2. The word document filled with solutions to all the problems.

3. The overall approach of the analysis in a presentation.
   1. Mention the problem statement and the analysis approach briefly
   2. Explain the results in business terms
   3. Include visualisations and summarise the most important results in the presentation

4. A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

Solution

□ Data cleaning and data manipulation.

1.1. Check and handle duplicate data.

2.2. Check and handle NA values and missing values.

3.3. Drop columns, if it contains large amount of missing values and not useful for the analysis.

4.4. Imputation of the values, if necessary.

5.5. Check and handle outliers in data.

□ EDA

1. Univariate data analysis: value count, distribution of variable etc.
2. 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 and recommendations.

Data Manipulation

- Total Number of Rows =37, Total Number of Columns =9240.

- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “and Update me on Supply”.

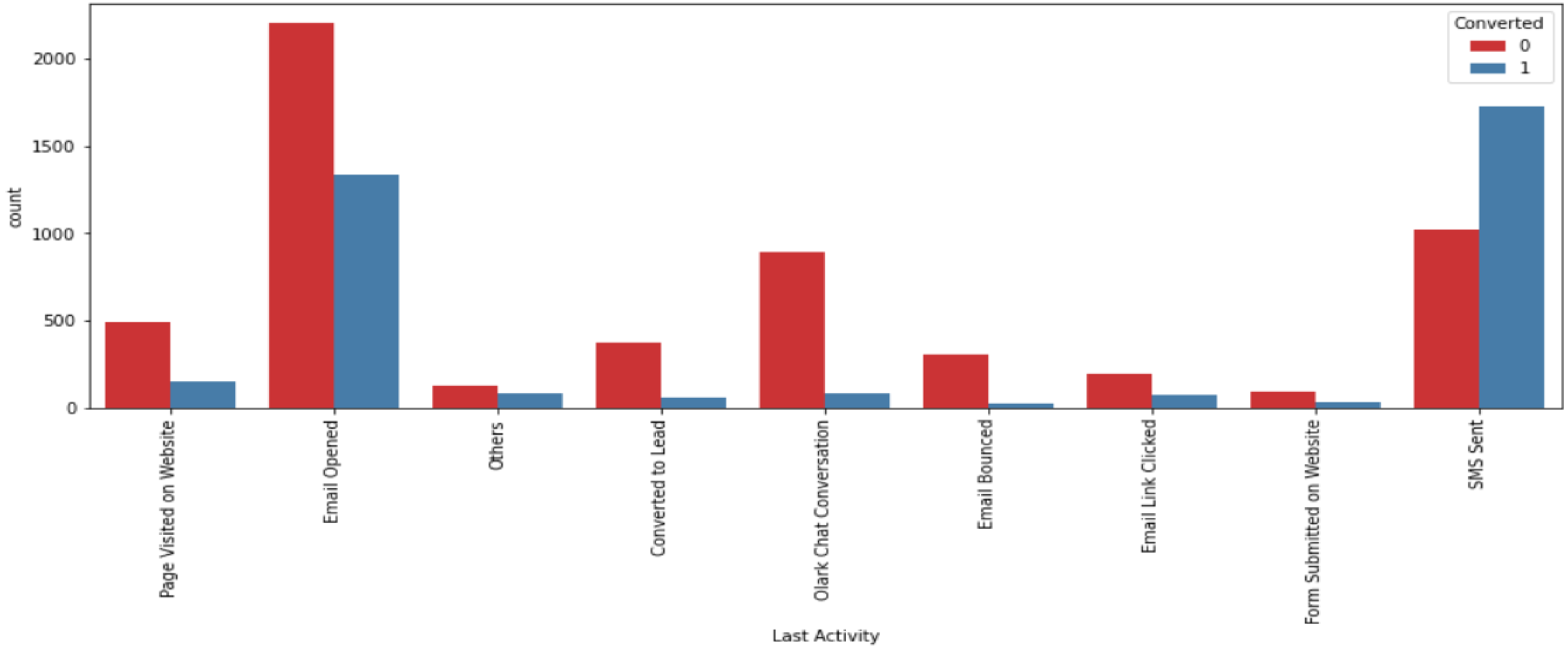
- “Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.

- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.

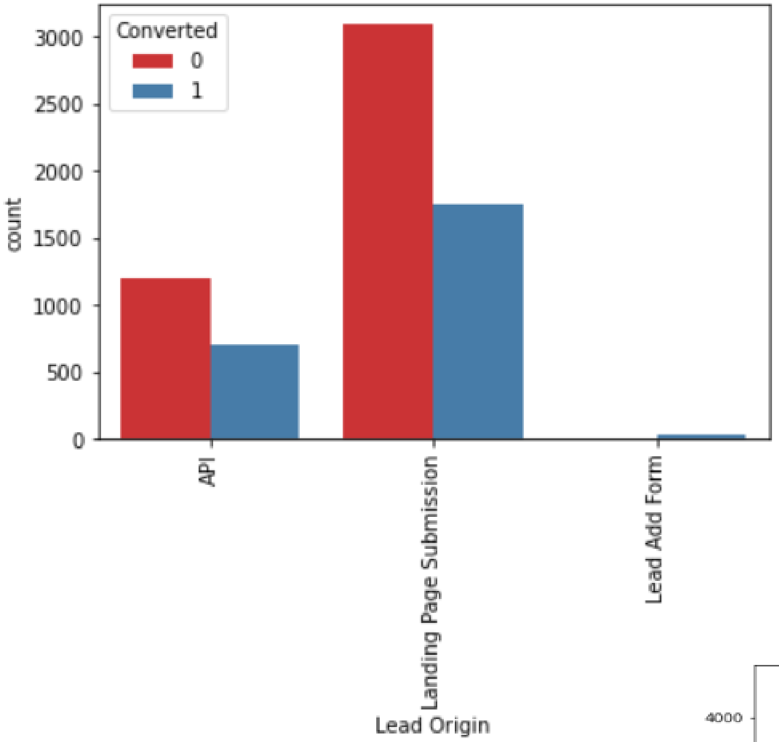
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.

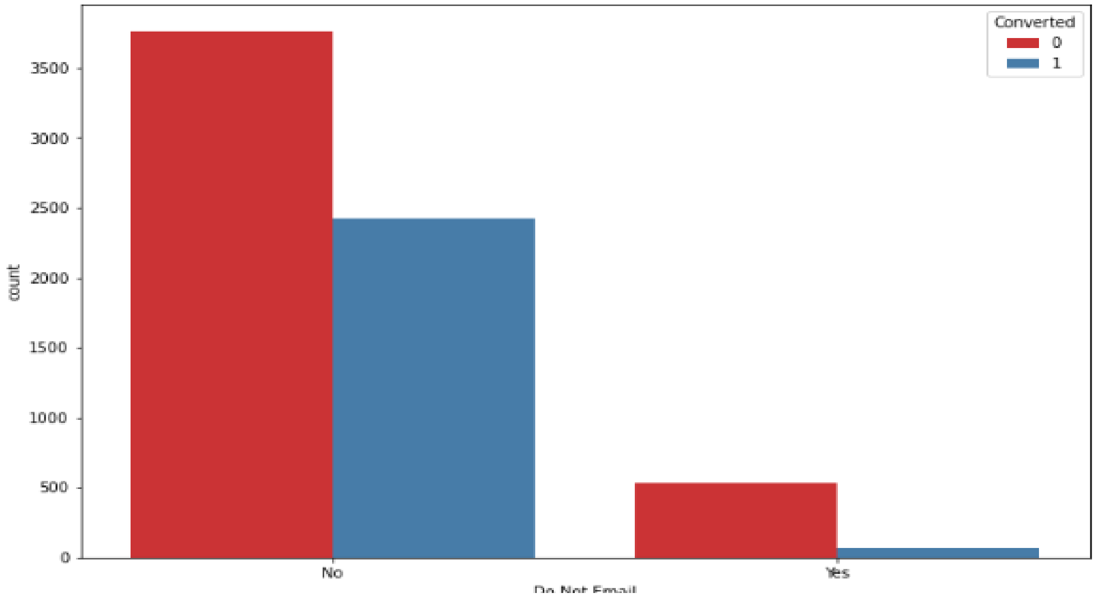
- Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

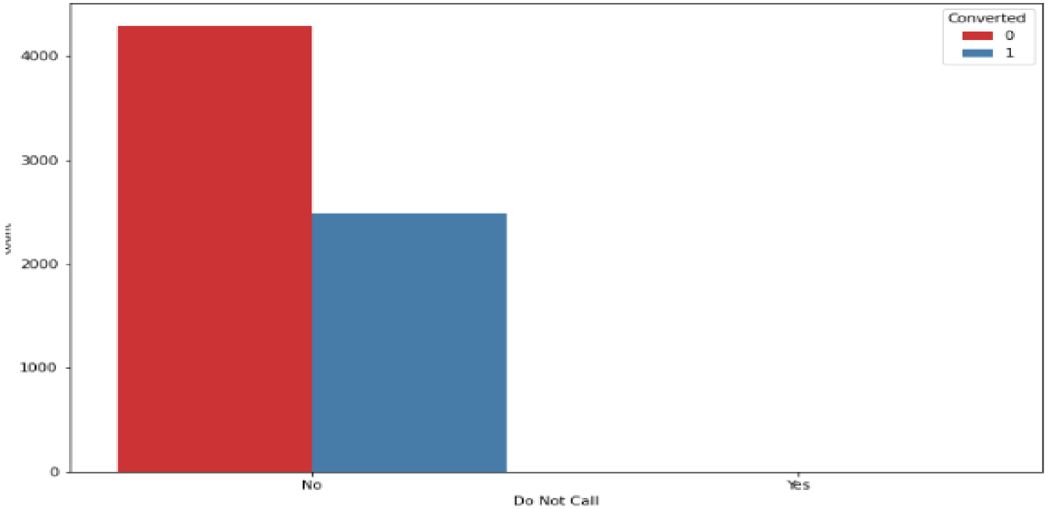
EDA



Converted
0
1
count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Others
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
SMS Sent
Last Activity


Converted
0
1
3000
2500
2000
1500
1000
500
0
count
API
Landing Page Submission
Lead Add Form
Lead Origin

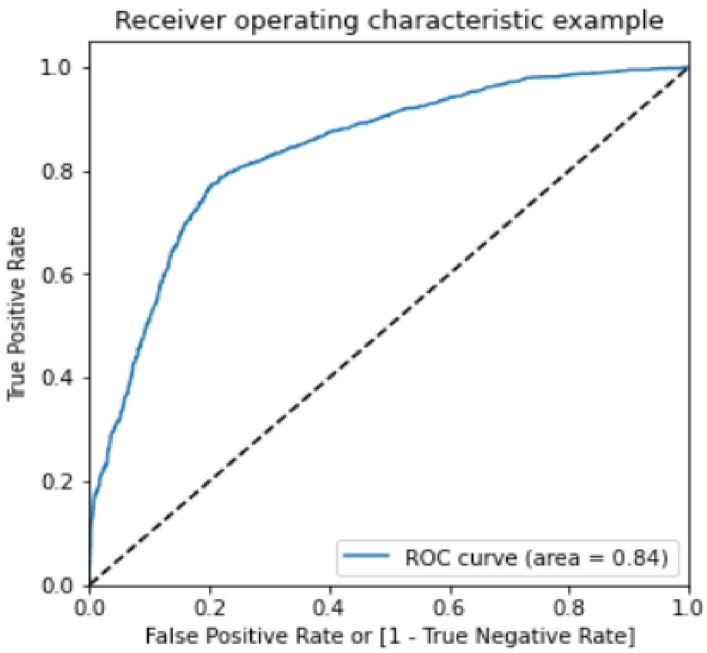

Converted
0
1
3500
3000
2500
2000
1500
1000
500
0
count
No
Yes
Do Not Email


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

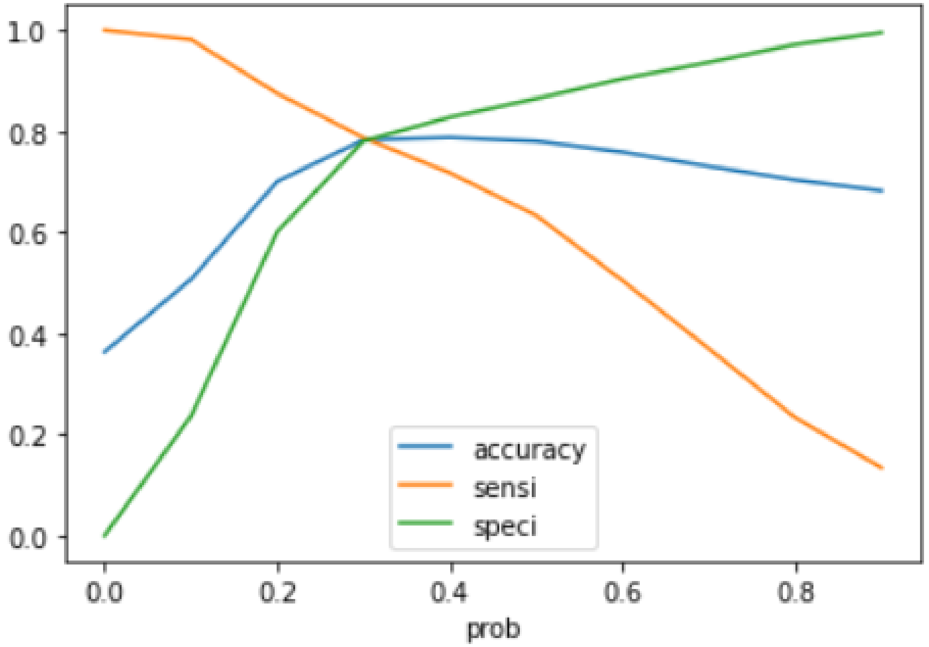
Model Building

- 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 15 variables as output.
- Building Model by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5.
- Predictions on test data set.
- Overall accuracy 81%.

ROC Curve



Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.84)



1.0
0.8
0.6
0.4
0.2
0.0
accuracy
sensi
speci
prob

- 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 0.35.

Conclusion

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

 The Total Time Spent on Website.

 Total number of Visits.

 When The Lead source was : a. Olark chat b. Wellingak Website

 When the Last Activity was : a. SMS b. Olark Chat Conversation

 When the lead origin is Lead add Form.

 When the current occupation was : a. Working Professionals b. Student c. Unemployed d. 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.