Lead Score Case Study

By Shivansh Ghildiyal, Jitendra Kumar Pradhan and Sugata Ranjan Das

Problem Statement:

An educational company X Education sells online courses to industry professionals and markets its courses on several websites.
When arriving at the website, the visitors may browse the courses, submit a form for the course, or watch some videos. These persons are categorized as leads when they fill out a form with their phone number or email address. Also, the business receives leads from earlier recommendations. Once these leads are obtained, sales team members begin calling, sending emails, etc. Some leads are converted during this procedure, but most are not. At X Education, the normal lead conversion rate is roughly 30%.

Business Goal:

The business wants to create a model in which give each lead is assigned with a lead score so that leads with higher lead scores have a better chance of converting, while leads with lower lead scores have a lesser chance of converting. The desired lead conversion rate has been estimated by the CEO is about 80%.

The objectives of the study are:

- Basic handing of the data, and deal with missing values with best measures.
- Check whether the data consists of outlier.
- Check the significant relationship between the target variable and other important variables by proper statistical analysis and data visualisation.
- Evaluate the model based on Sensitivity and Specificity.

Procedure to deal the data:

The primary data set is the “lead.csv”, where information regarding the visitors on the website, is given.

First the shape and size of the data is understood. After that the columns of with more than 40% of missing values are omitted.

After the missing value treatment the outlier check is perfumed and after taking necessary steps, the data is ready for the analysis.

The final data is categorized in trail and test data set in 7:3 ratio.

Using Recursive Feature Elimination a logistic model is fit on the train data set with acceptable accuracy measures. Now, that fitted model is used to predict the test data set.

34
34
341880

Missing value dealing:

The original data consists of 9240 rows and 37 columns with 341880 observations. The variables are in the form of float, integer and object. There are 5 columns with more than 40% missing values. After removing those missing values, 32 columns are remaining.

There are still a few columns, which are irrelevant for the study. Such that, “City” and “Country”. It is needed to drop these variables.

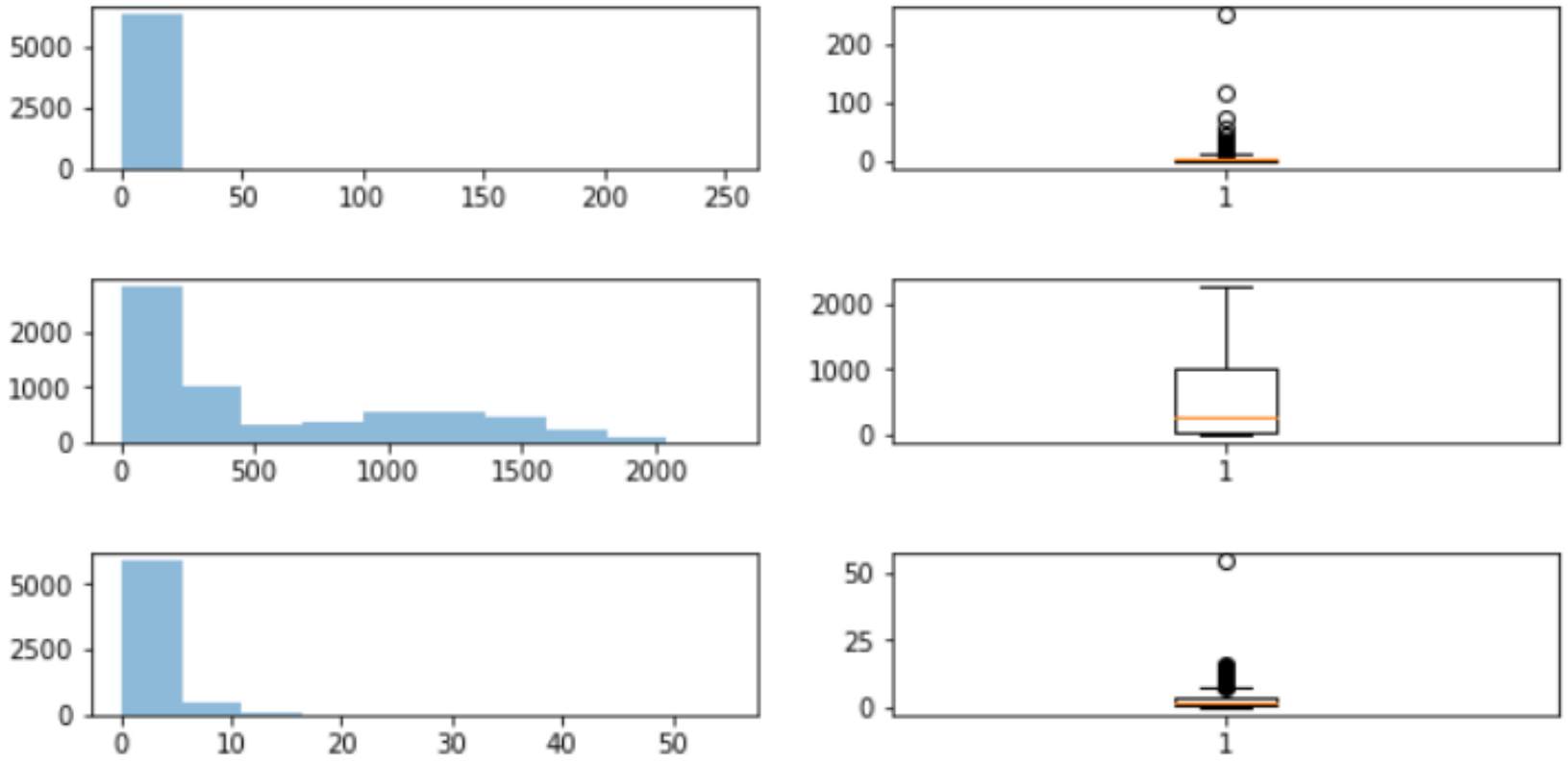
After removing these irrelevant columns, there are 30 columns to deal with.

Data imbalance:

- The label "select" under some variables implies that the subject had not selected this option while filling the form. Thus, "select" is also considered as null value. The variables "Lead Profile", "How did you hear about X Education" and "Specialization“ have the label "select". The number of "select"s are very high for "Lead Profile", and "How did you hear about X Education". And, there are a few variables with majority of values are marked as "no". Better to drop these variables.
- The dummy variables are created for categorical variables with multiple levels and the repeated variables are removed.

Outlier check:

- The quantitative variables “TotalVisits”, “Total Time Spent on Website”, “Page Views Per Visit” consist of outliers.



The distributions of the continuous variables are skewed and does not follows normal distribution. Moreover, the variables "TotalVisits" and "Page Views Per Visit" are having outliers.

Any observation is higher than [Q3 + 1.5 * IQR] or lower than [Q1 - 1.5 * IQR] is considered outliers.

After dealing with outliers the final dataset is lead1 consists of 5987 rows and 75 columns with 449025 observations.

The response variable is “Converted”.

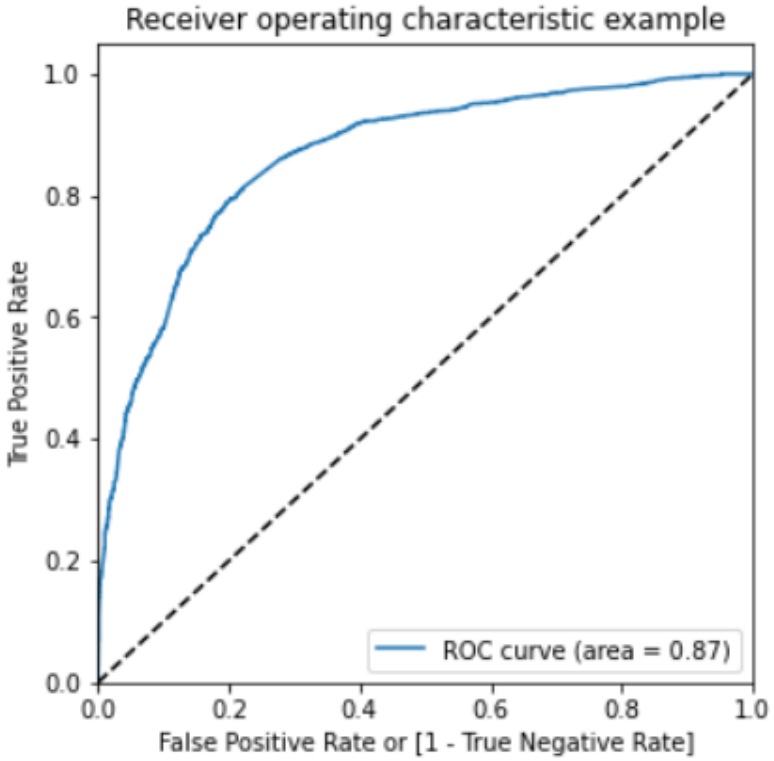
- In the final logistic model, the variables those impact the target variable “Converted” are **Total Time Spent on Website**, **Lead Origin_Lead Add Form**, **Lead Source_Olark Chat**, **Lead Source_Welingak Website**, **Do Not Email_Yes**, **Last Activity_Email Bounced**, **Last Activity_Had a Phone Conversation**, **Last Activity_Olark Chat Conversation**, **Last Activity_SMS Sent**, **What is your current occupation_Student**, **What is your current occupation_Unemployed**, and **Last Notable Activity_Unreachable**.

Model evaluation on train data (based on accuracy, sensitivity, and specificity):

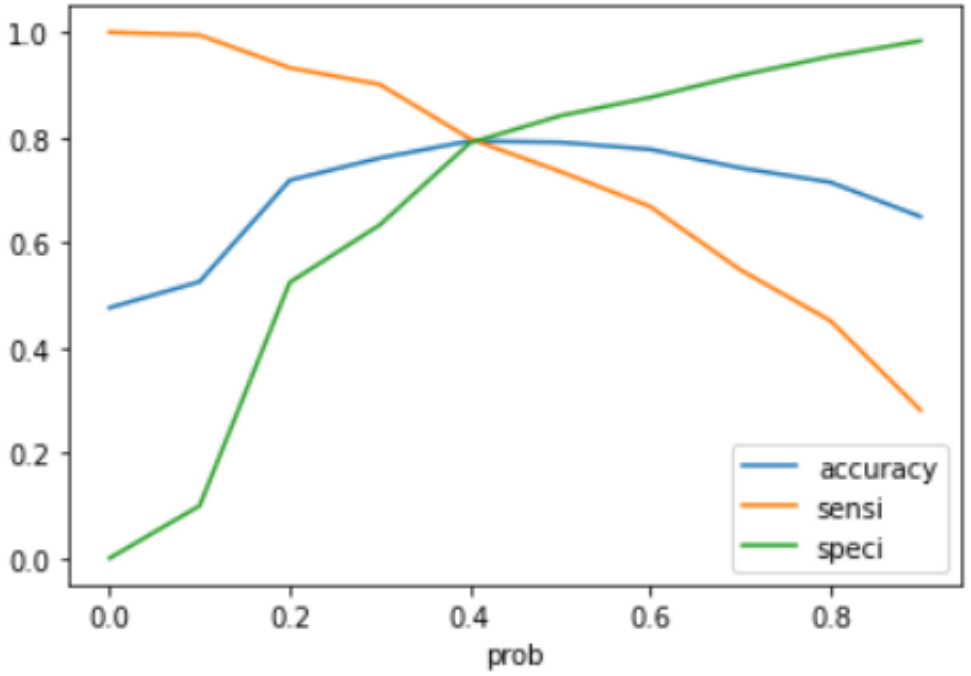
Accuracy: 80%
Sensitivity: 79%
Specificity: 80%
False Positive rate: 20%
Positive Predictive rate: 74%
Negative Predictive rate: 81%

Confusion Matrix		
Actual/Predicted	Negative	Positive
Nrgative	1762	433
Positive	421	1574

ROC curve:



From the below graph the optimal cut-off point is 0.42

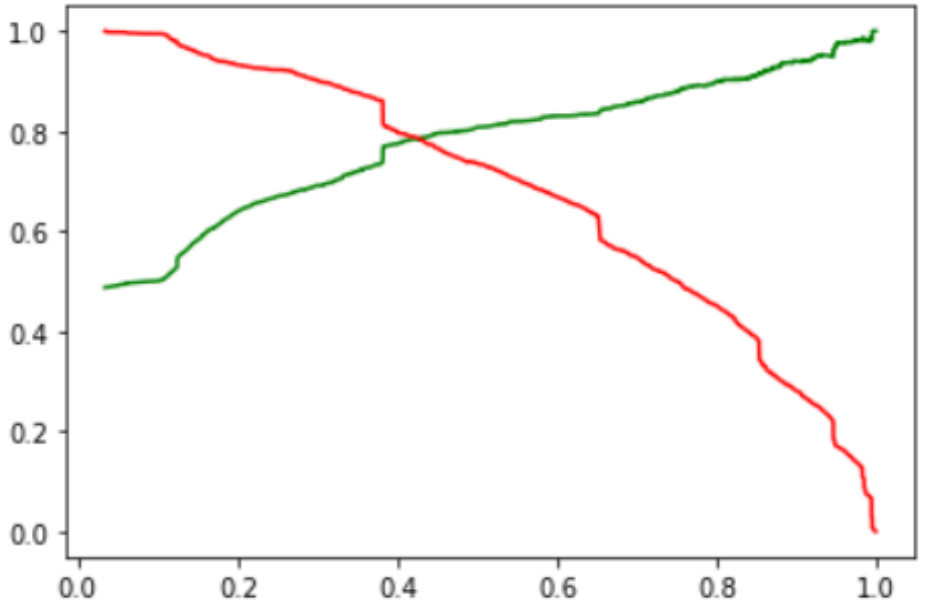


Model evaluation on train data (based on precision and recall):

Precision: 81%
Recall: 73%

Confusion Matrix		
Actual/Predicted	Negative	Positive
Negative	1846	349
Positive	528	1467

From the below graph the optimal cut-off point is 0.41



Model evaluation on test data:

Accuracy: 77%
Sensitivity: 76%
Specificity: 77%

Confusion Matrix		
Actual/Predicted	Negative	Positive
Negative	716	217
Positive	189	675

Final conclusion:

From the analysis, the top three variables in the model which contribute most towards the probability of a lead getting converted are, “Total Time Spent on Website”, “What is your current occupation_Student”, and “What is your current occupation_Unemployed”.

The top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion are, “Last Activity_SMS Sent”, “Lead Source_Olark Chat” and “Lead Origin_Lead Add Form”.

It is obtained that the lead origin as well as lead source are significant dummy variables in order to increase the probability of lead conversion. In addition, the X-edication team should target those people who spent higher time on the website. Among these people they need to focus mainly only those people who are not students and who are unemployed or employed, but want to have more business knowledge. The interns should reach them more competitively and explain them how good the X-education is compared to other institutions.