Lead Scoring Case Study

Submitted by,
Ankit Tomar, Swamini Lunge, Rohit Tehlan

Problem Statement

Industry professionals can purchase online courses from X Education, an education firm. Many experts who are interested in the courses visit their website and look through the offerings on any given day. The business advertises its classes via a number of websites and search engines, including Google. Upon accessing the website, these individuals may peruse the available courses, complete the course registration form, or view some videos. These folks are categorized as leads when they complete a form with their phone number or email address. Additionally, the business receives leads from previous recommendations. After obtaining these leads, sales team members begin calling, emailing, and so on. Some of the leads convert during this process, whereas most do not.

The typical lead conversion rate at X education is around 30%. Few leads emerge from the bottom as paying clients, despite the fact that many are created in the first stage (top). To increase lead conversion in the intermediate stage, you must properly nurture your potential leads (i.e., educate them about the product, stay in continual communication, etc.). You have been assigned by X Education to assist them in identifying the most promising prospects—that is, the leads with the highest likelihood of becoming paying clients. The business needs you to create a model in which you give each lead a score so that customers with higher lead scores have a better likelihood of converting, while customers with lower lead scores have a reduced chance of doing so. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Goals of the Case Study

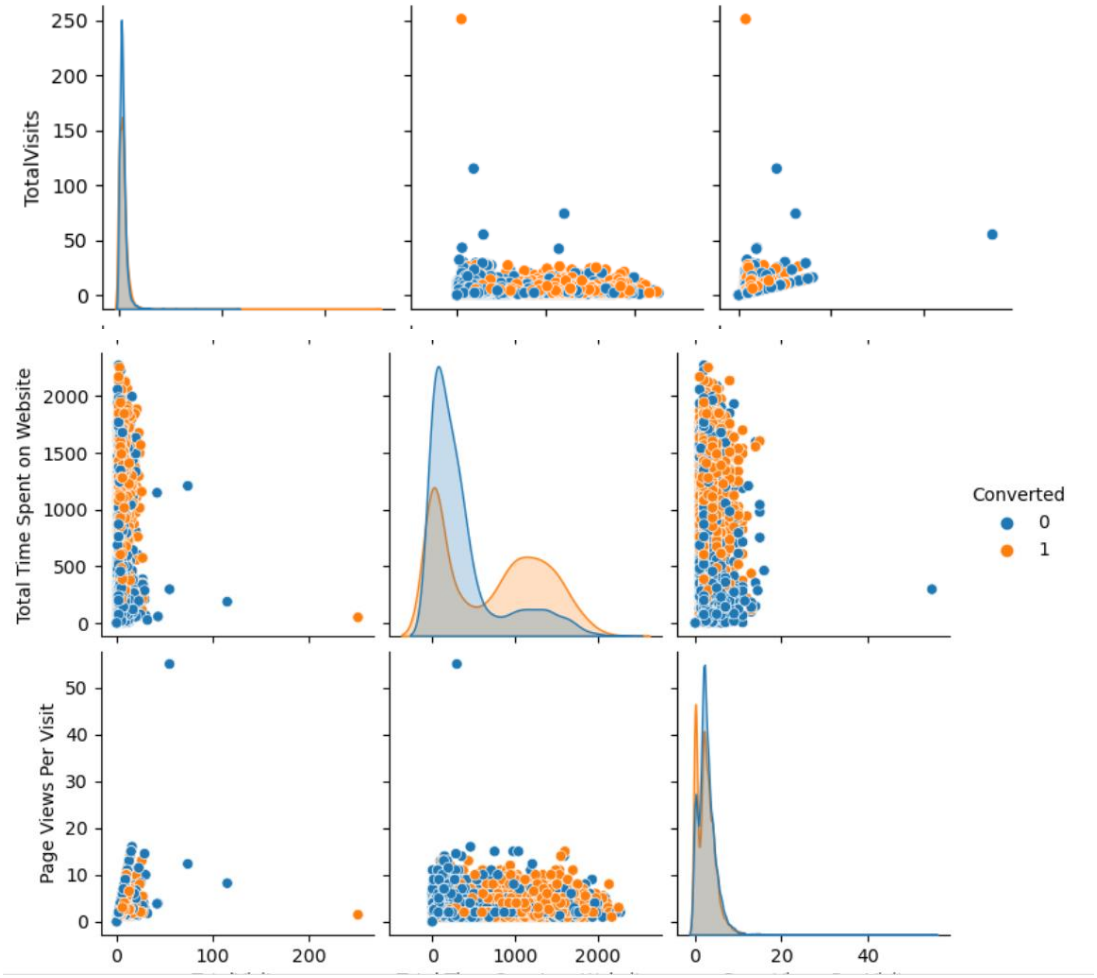
- To build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

- To adjust to if the company's requirement changes in the future so you will need to handle these as well.

The steps involved

- Reading and understanding the data
- Data cleaning
- Prepare the data for Model Building
- Model Building
- Model Evaluation
- Making Predictions on the Test Set

Plots for identifying impacting factors



Inferences:

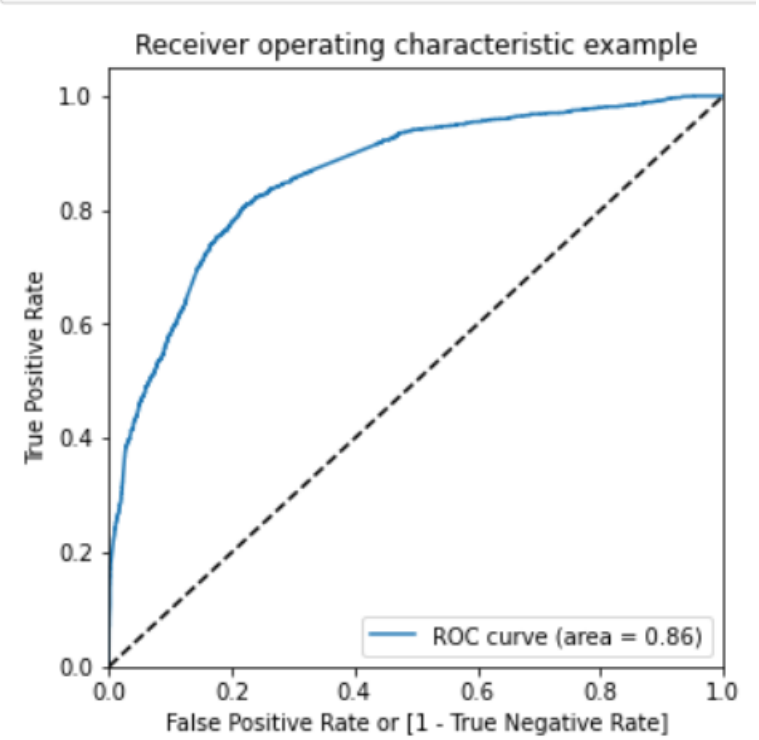
"Total time spent on website" has higher impact in converting leads.

Initial Observations

Train data:

- Accuracy = 78.86%
- Sensitivity = 73.94 %
- Specificity = 83.43%

Graphs



The area under the curve of the ROC is 0.86 which is quite good. So we seem to have a good model. Check the sensitivity and specificity tradeoff to find the optimal cutoff point, gives below numbers:

Accuracy = 79.08%
Sensitivity = 79.33 %
Specificity = 78.84%

This cut off point seems good to go!

Final Observations on test set

- Overall Accuracy = 78.66%
- Precision = 78.28 %
- Recall = 76.74%

Suggestions

Don’ts:

- Do not focus on unemployed leads. They might not have a budget to spend on the course.
- Do not focus on students, since they are already studying and would not be willing to enroll into a course specially designed for working professionals, so early in the tenure.
- Do not focus on those who have already enrolled for any free course, as they might looking to spend or invest further on courses.

Dos:

- Total Time Spent on Website: Sales team can focus on such leads
- Lead Source Reference: If the source of the lead is a Reference, then there is a higher probability that the lead would convert.