Lead Scoring Case Study

- Kanagaraj Rajendran
- Sanjay Sharma
- Sanjay Kumar Singh

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Eventhough X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. 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.

There are a lot of leads generated in the initial stage but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

X Education wants select the promising leads by data analysing, i.e. the leads that are most likely to convert into paying customers. The company requires 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

1. 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. 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. The model can able adjust to if the company's requirement changes in the future so you will need to handle these as well.

Approach

Steps followed for the Analysis are as follows:

- ➢Data read and inspection
- ➢Clean and prepare the data
- ➢Exploratory Data Analysis & Data Visualization
- ➢Feature Scaling
- ➢Splitting the data into Test and Train dataset
- ➢Building a logistic Regression model and calculate Lead Score
- ➢Evaluating the model by using different metrics -Specificity and Sensitivity or Precision and Recall.
- ➢Applying the best model in Test data based on the Sensitivity and Specificity Metrics.

Data Reading & Inspection

- ➢Read the CSV file using pd.read
- ➢Check the shape of the data
- ➢Check the columns names and understand the column
- ➢Check the data types for all the columns
- ➢Describe the data

Data Preparation

- Find the Null values of each column
- Removed those columns have more than 30% of missing values
- Removed those columns have only one value counts to avoid infusion on the prediction
- Removed those rows having higher missing values

Data Visualization

- View pair plot for understand the co-relation with each columns
- Found few variables are positively co-related

Scaling and Splitting Train and Test data

- Create dummies for categorical variable
- Splitting the data into Train and Test (70 % Train, 30% Test data)
- Align the numeric variable data into standard MinMax scalar

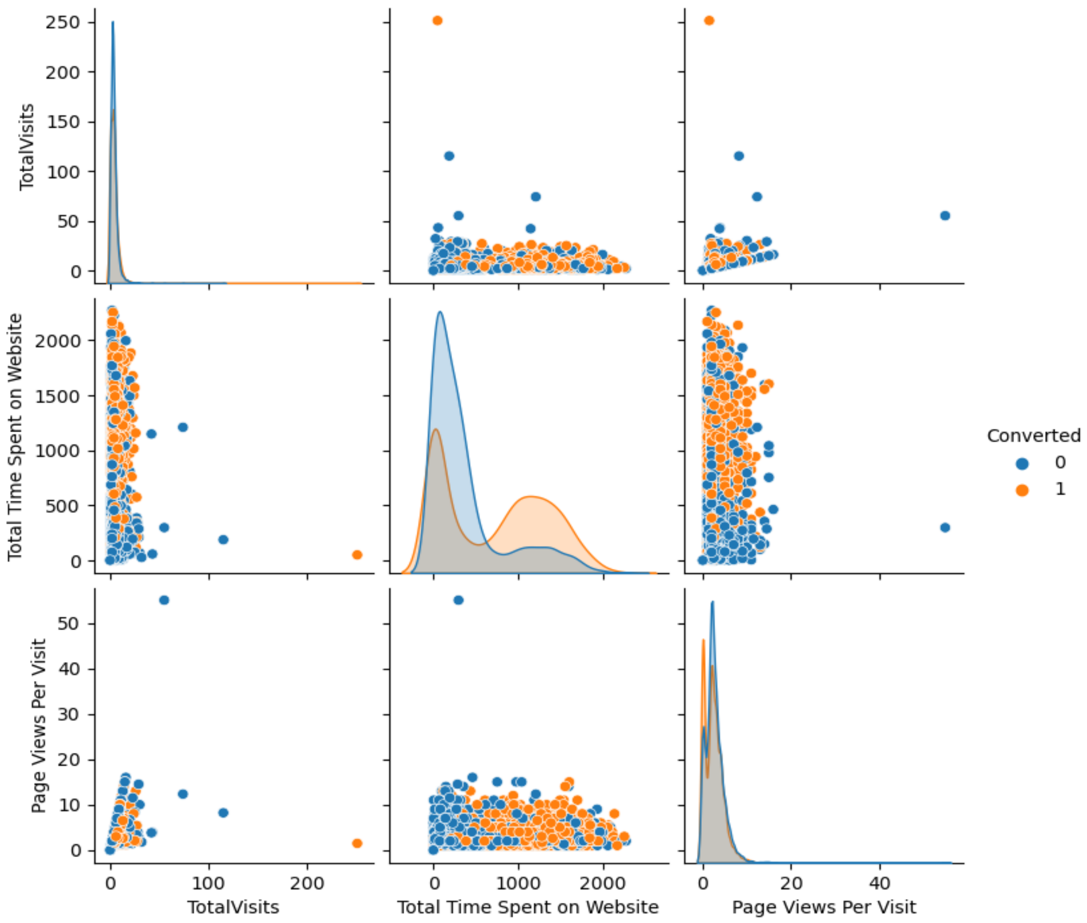
Model Building

- Feature selection using RFE
- Assessing the model with statsModels using Logistics Regression
- Find the VIF and compare the P-Value for select the best feature
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall to evaluate the model

Result

- Find the lead score and check if target final predication amounts to 80% conversion rate.
- Evaluate the final predication on the test set using cutoff threshold from sensitivity and specificity metrics

Exploratory Data Analysis



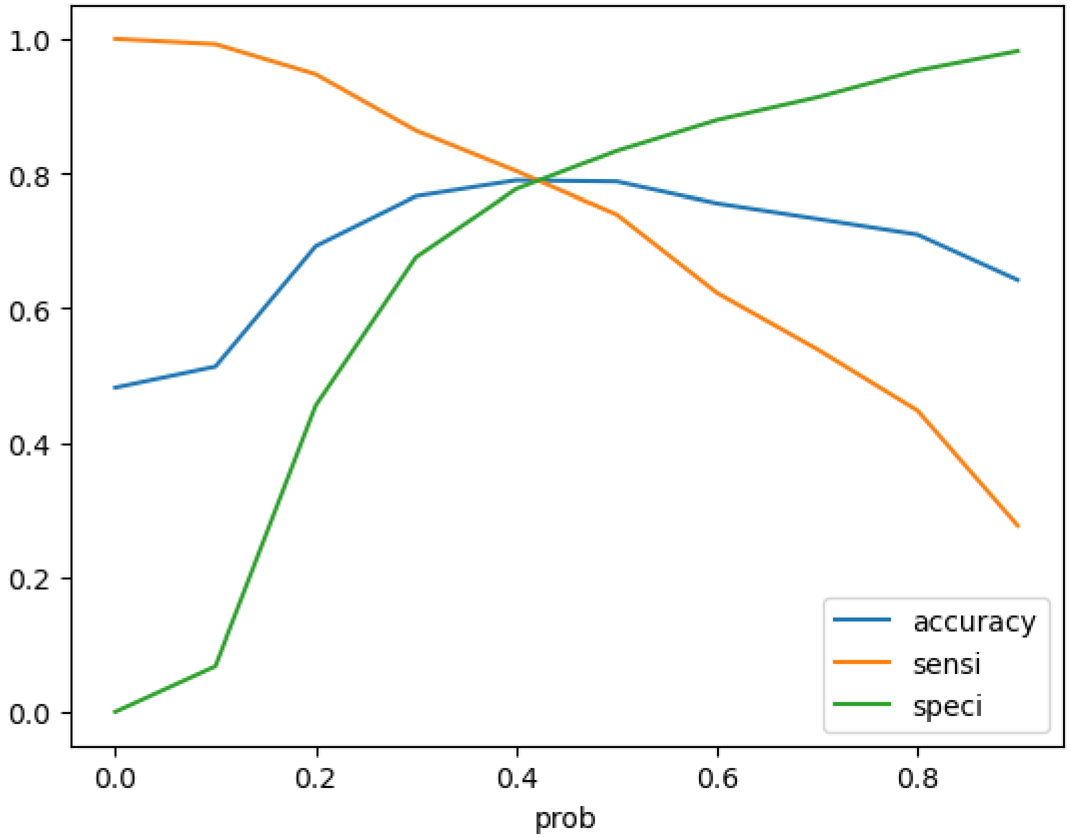
- ➢Based on the visual , we could see that the Total visit and Page Views per visit has positively co-related

- ➢Let's check further for the categorical variable. To analysis the categorical column first will create dummies

Variables Impacting the Conversation Rate

- TotalVisits
- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Lead Source_Olark Chat
- Lead Source_Welingak Website
- Last Activity_Email Bounced
- Last Activity_Had a Phone Conversation
- Last Activity_SMS Sent
- What is your current occupation_Student
- What is your current occupation_Unemployed
- Last Notable Activity_Unreachable

Model Evaluation – Sensitivity and Specificity on Train Data Set



- The graph shows the cut off of 0.42 based on Accuracy, Sensitivity and Specificity
- Accuracy: 79 %
- Sensitivity: 78%
- Specificity: 79%

Confusion Matrix

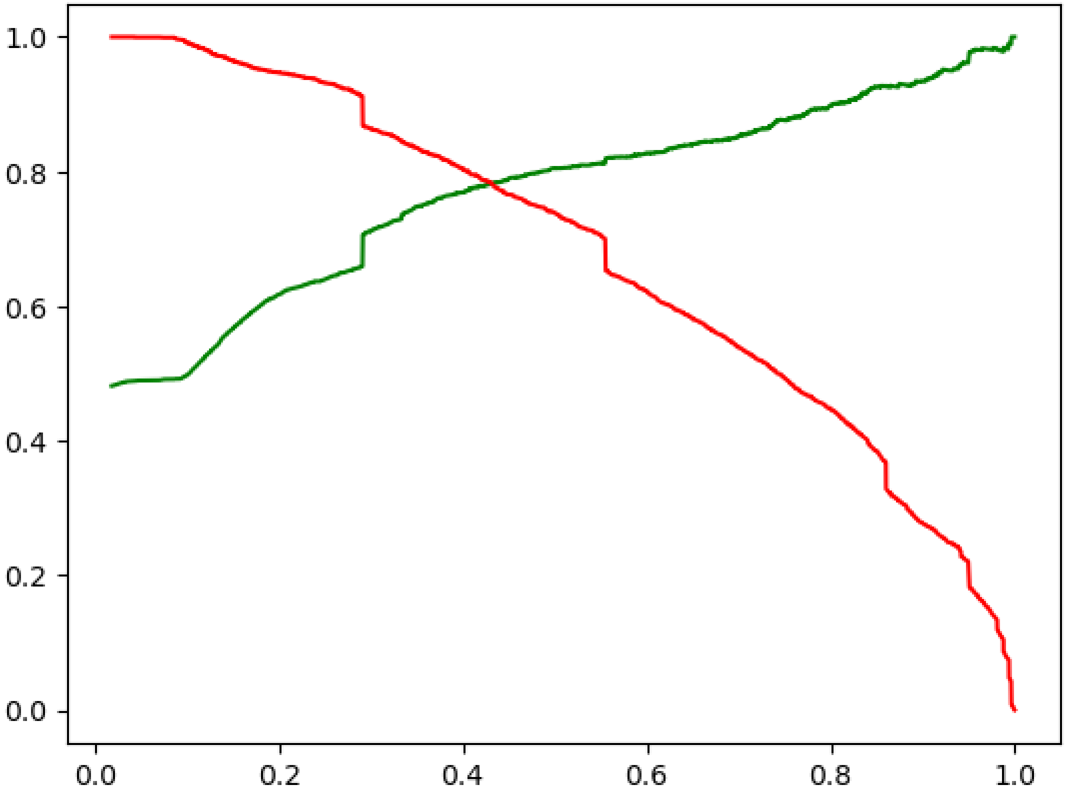
1830	482
454	1695

Model Evaluation – Sensitivity and Specificity on Test Data Set

- Accuracy: 79 %
- Sensitivity: 78%
- Specificity: 80%
- Confusion Matrix

792	204
201	715

Model Evaluation – Precision and Recall on Train Data Set



- The graph shows the cut off of 0.42 based on Precision and Recall
- Precision: 78%
- Recall: 79%
- Confusion Matrix

1830	482
454	1695

Conclusion

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction. –
- Accuracy, Sensitivity and Specificity values of test set are around 79%, 78% and 80% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated shows the conversion rate on the final predicted model is around 80% (in train set) and 79% in test set

Suggestions & Recommendations

- Target leads that spend a lot of time on X-Education Website
- Target leads that repeatedly visit the site. However they might be repeatedly visiting to compare courses from the other sites, as the number of visits might be for that reason. So the interns should be a bit more aggressive and should ensure competitive points where X-Education is better, are strongly highlighted.
- Target leads that have come through References as they have a higher probability of converting
- The sales team can conduct the social media campaign with post of course updates, webinars, Instructor bio, the successful learner feedback, to keep highlight their platform
- Search engine optimization is one of the task that sales team perform to keep their website to be in the top when learns search with different key words
- Question and feedback form to the learns already enrolled in between the course, will helpful to get the mindset of their choosing course and their expectation
- The last one is speed up the platform run time
- Do not call the housewives, students enrolled in long term courses.