Lead Score Case Study

Submitted BY

Bijay Khanal

Lead Score Case Study for X Education

Problem Statement :

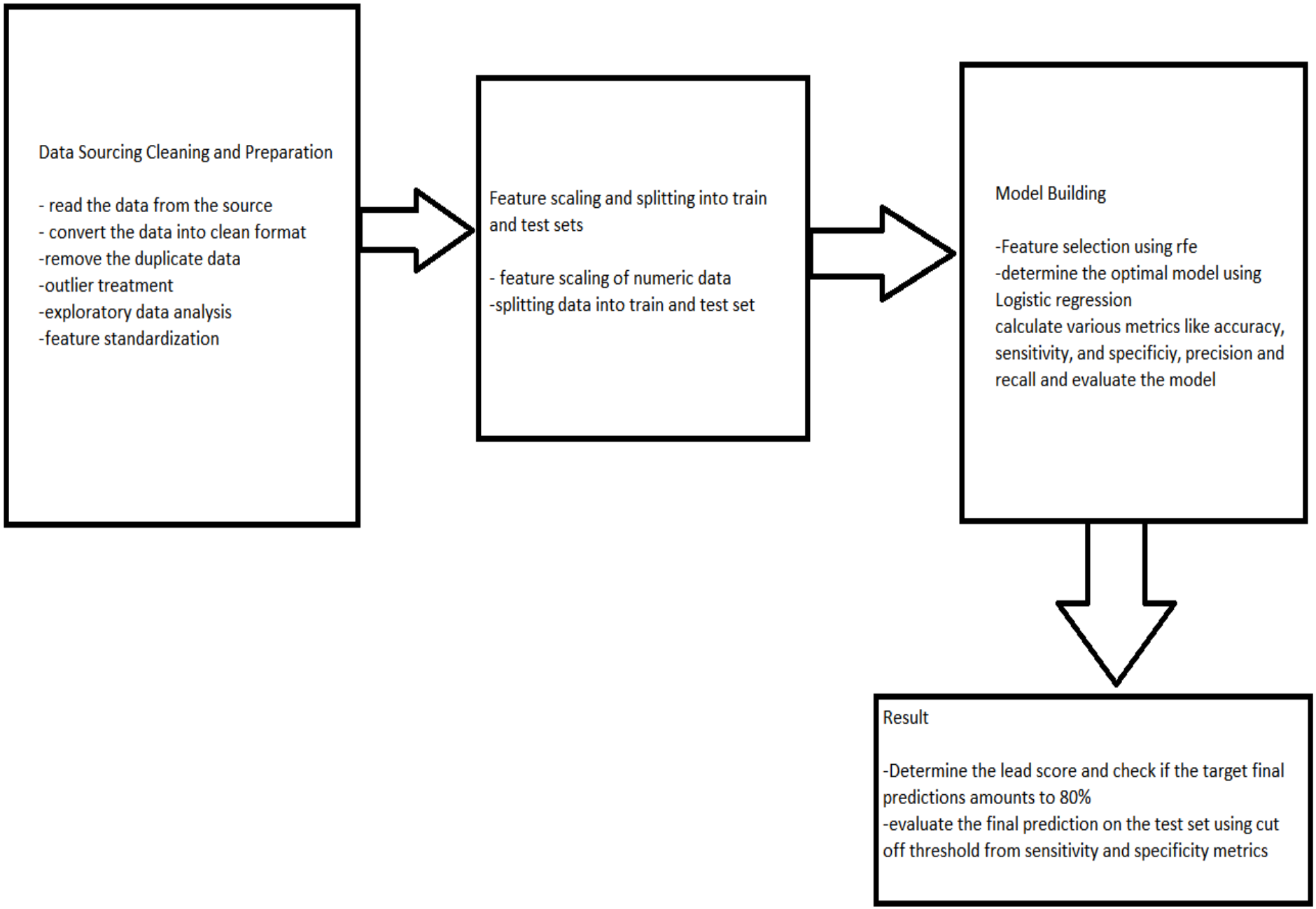
- An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.
- Once the people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When 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%.

Business Goal:

- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with 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%.

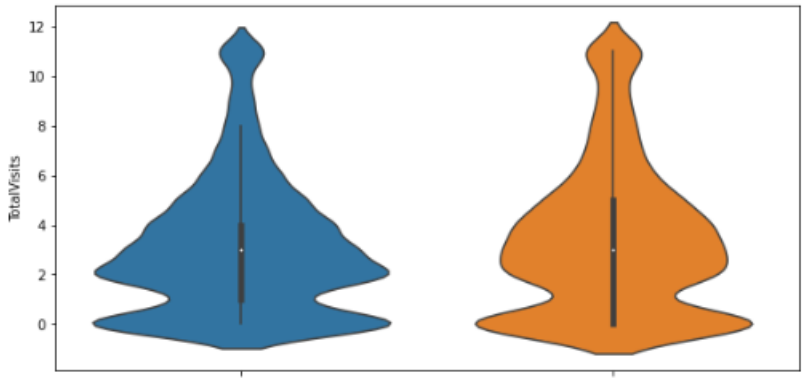
Strategy

- Source the data for analysis
- Clean and prepare the data
- Exploratory Data Analysis.
- 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.

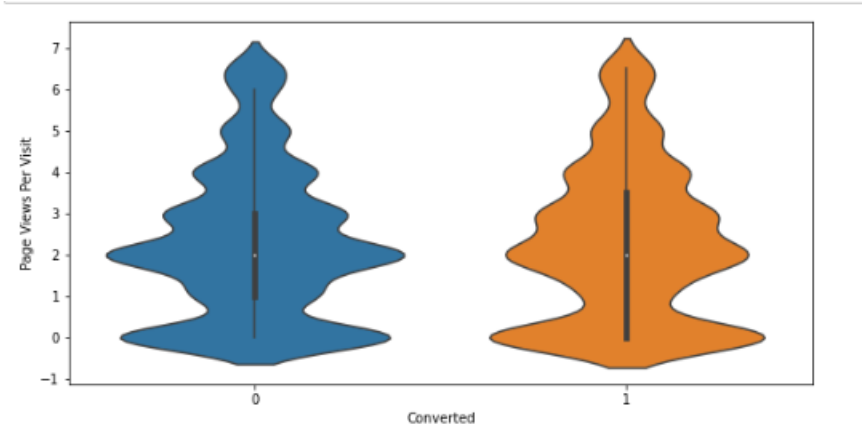

Data Sourcing Cleaning and Preparation
- read the data from the source
- convert the data into clean format
-remove the duplicate data
-outlier treatment
-exploratory data analysis
-feature standardization
Feature scaling and splitting into train and test sets
- feature scaling of numeric data
-splitting data into train and test set
Model Building
-Feature selection using rfe
-determine the optimal model using Logistic regression
calculate various metrics like accuracy, sensitivity, and specificiy, precision and recall and evaluate the model
Result
-Determine the lead score and check if the target final predictions amounts to 80%
-evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

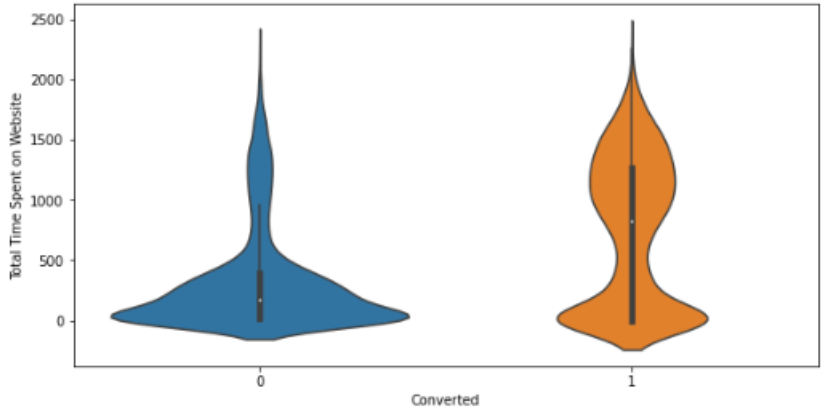
Exploratory data analaysis

The conversion rates for Total Visits Total time spent on website and Page Views per visit were higher.

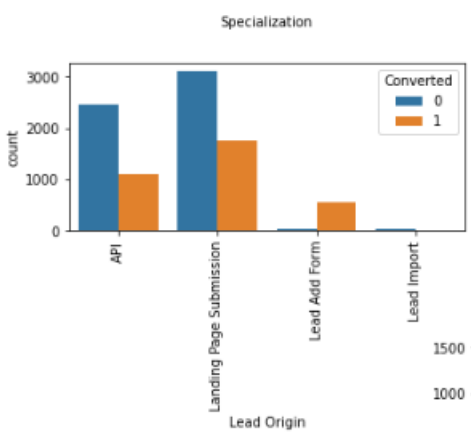


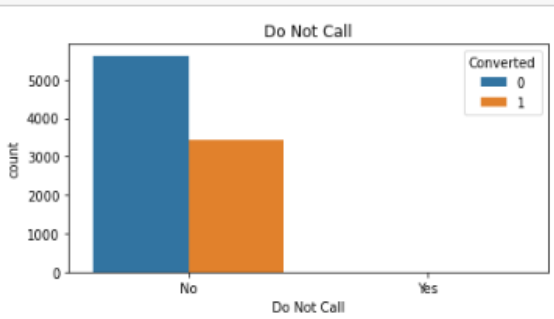
The conversion rates for the TotalVisits Time spent on website and Page views Per visit are high

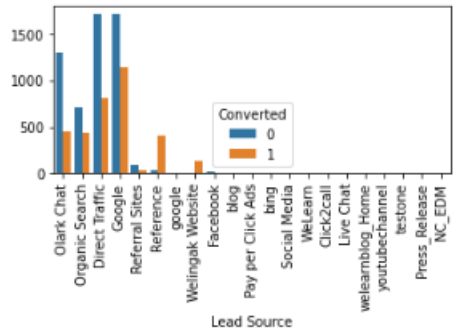




- In lead origin Landing page submission the highest conversion occurred
- In emails sent and calls made also highest conversion occurred
- Lead source Google has maximum conversion



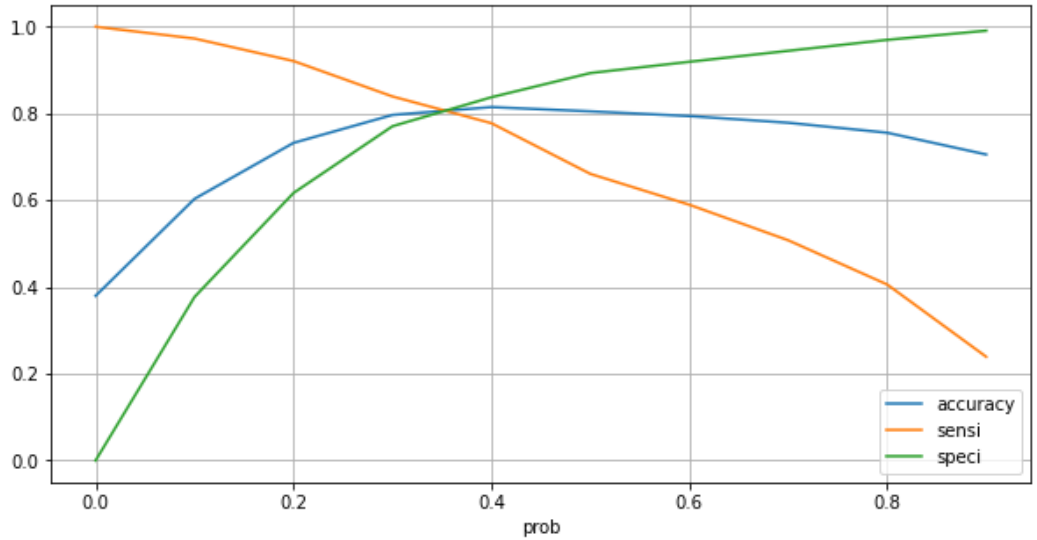




Variables Impacting the Conversion Rate

- Do not Email
- Total Time Spent on Website
- Lead Origin_Landing Page Submission
- Lead Origin_Lead Add Form
- Lead Source_Olark Chat
- Lead Source_Reference
- Lead Source_Welingak Website
- Last Activity_Email Opened
- Last Activity_Olark Chat Conversation
- Last Activity_Other_Activity
- Last Activity_SMS Sent
- Last Activity_Unreachable
- Last Activity_Unsubscribed
- Specialization_Others
- What is your current occupation_Housewife
- What is your current occupation_Student
- What is your current occupation_Unemployed
- What is your current occupation_Working Professional

Model Evaluation -Sensitivity and Specificity on Train Data Set



This graph illustrates an optimal cutoff of 0.34 depending on the accuracy, sensitivity and specificity

The confusion matrix is

[[3607, 896],

[508, 2248]]

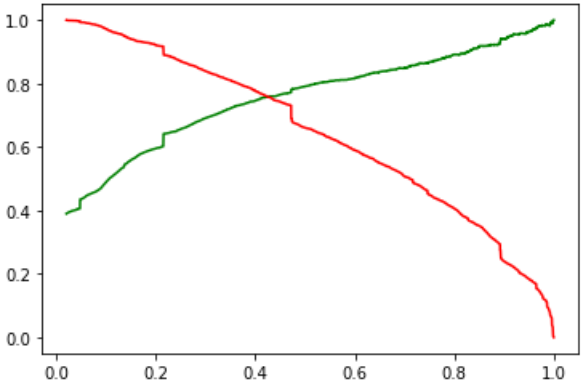
Train Data Accuracy :80.49 %

Train Data Sensitivity :81.57 %

Train Data Specificity :80.1 %

Train Data F1 Score :0.72

Model Evaluation-Precision and Recall on Train Dataset



Precision : 79.10%
Recall : 66.07%

Model Evaluation –Sensitivity and Specificity on Test Dataset

Confusion matrix is
[[928, 208],
 [155, 524]]

- Test Data Accuracy : 80.0 %
- Test Data Sensitivity : 81.57 %
- Test Data Specificity : 80.1 %
- Test Data F1 Score : 0.74
- Precision : 71.58%
- Recall: 77.17%

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 80%, 81% and 80% which are approximately closer to the respective values calculated using trained set.
- With insights from data, Team will be able to design a model to share lead Score , which will help sales team to target hot leads & convert them into Customers. Table UI is made simple to make sure teams works efficenctly & improves the sales.
- Based on the absolute coefficient values the top 3 variables are:

  Lead Origin_Lead Add Form

  Lead Source_Welingak Website

  Last Activity_Other_Activity Hence overall this model seems to be good.