Lead Score Case Study

Submitted By:

Rahul Raj

Shivangi Mishra

Problem Statement :

▶ X Education sells online courses to industry professionals. 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%.
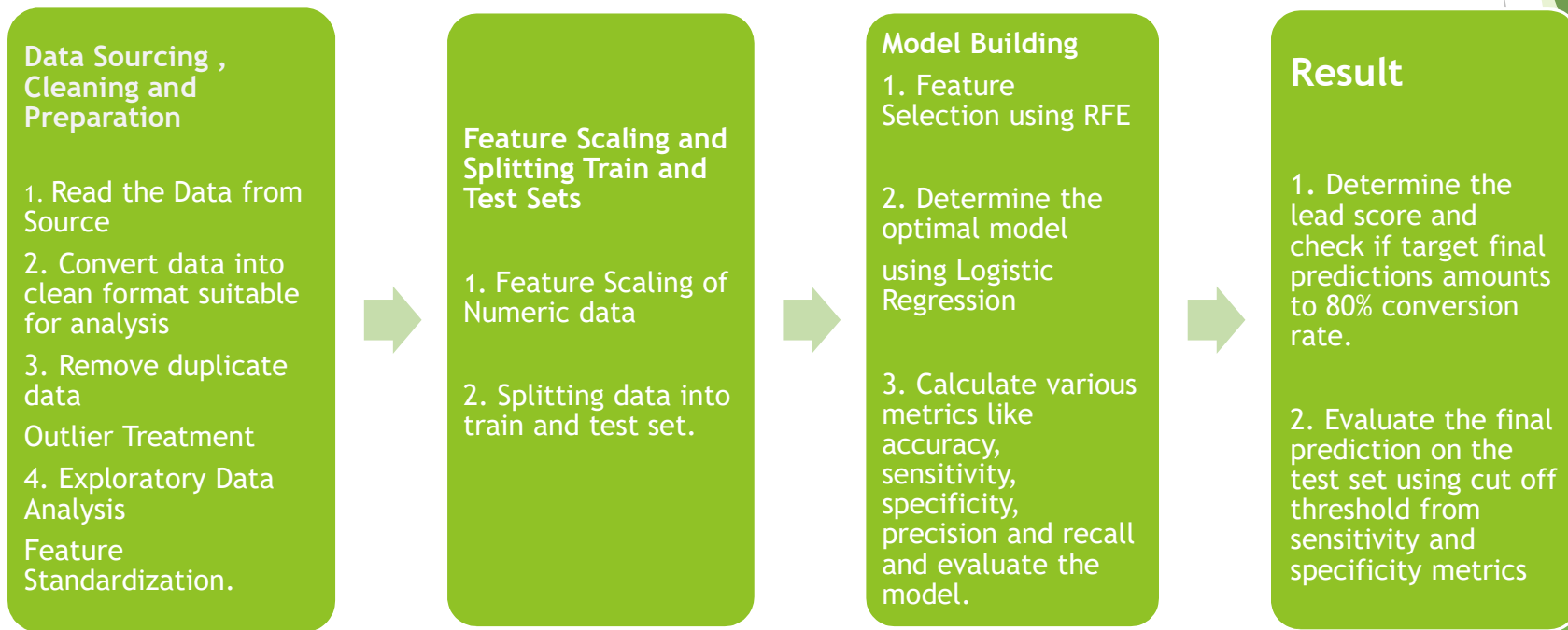
Business Goal:

▶ 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.

Problem solving methodology:

Data Sourcing, Cleaning and Preparation

1. Read the Data from Source
2. Convert data into clean format suitable for analysis
3. Remove duplicate data

Outlier Treatment

4. Exploratory Data Analysis

Feature Standardization.

Feature Scaling and Splitting Train and Test Sets

1. Feature Scaling of Numeric data
2. Splitting data into train and test set.

Model Building

1. Feature Selection using RFE
2. Determine the optimal model using Logistic Regression
3. Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

Result

1. Determine the lead score and check if target final predictions amounts to 80% conversion rate.
2. Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics
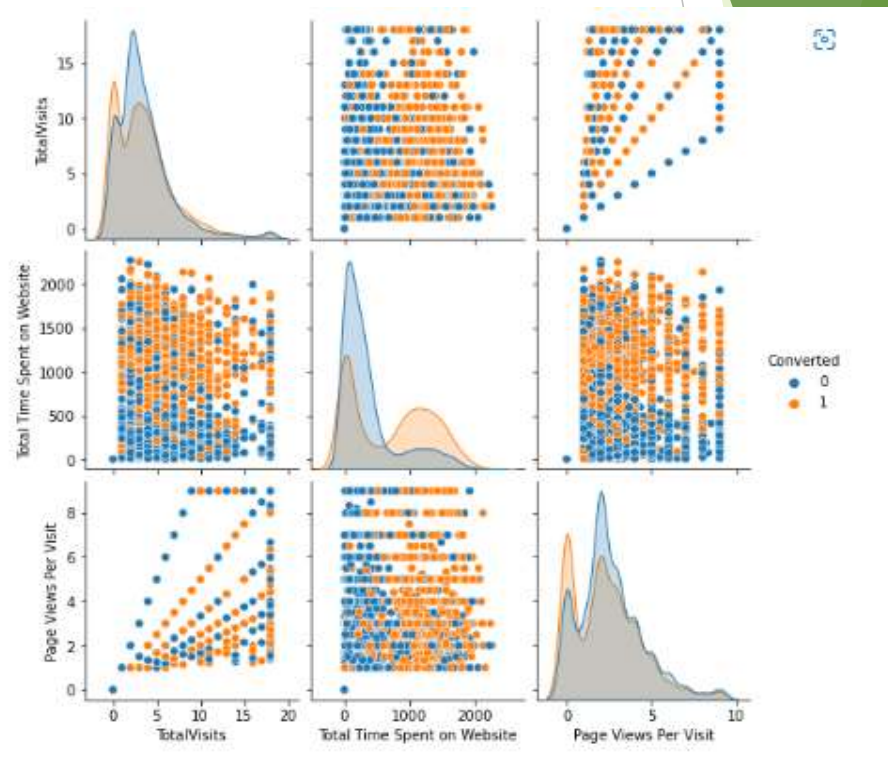
Exploratory Data Analysis

Balance of converted vs non converted in the cleaned data frame
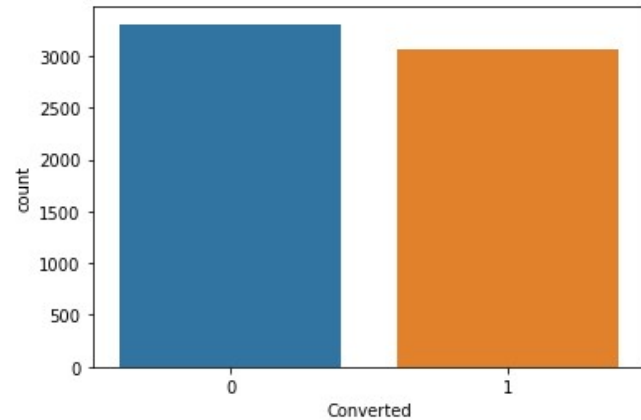
This is a fairly balanced data frame to go about - since the target variable is balanced at 52 - 48 ratio

Variables Impacting the Conversion Rate:

- Do Not Email
- Total Visits
- Total Time Spent On Website
- Lead Origin – Lead Page Submission
- Lead Origin – Lead Add Form
- Lead Source - Olark Chat
- Last Source – Welingak Website
- Last Activity – Email Bounced
- Last Activity – Not Sure
- Last Activity – Olark Chat Conversation
- Last Activity – SMS Sent
- Current Occupation – No Information
- Current Occupation – Working Professional
- Last Notable Activity – Had a Phone Conversation
- Last Notable Activity - Unreachable

Model Evaluation – Sensitivity and Specificity on Train Data Set:

▶ Considering optimal cut off 0.5, ROC curve shows as below:
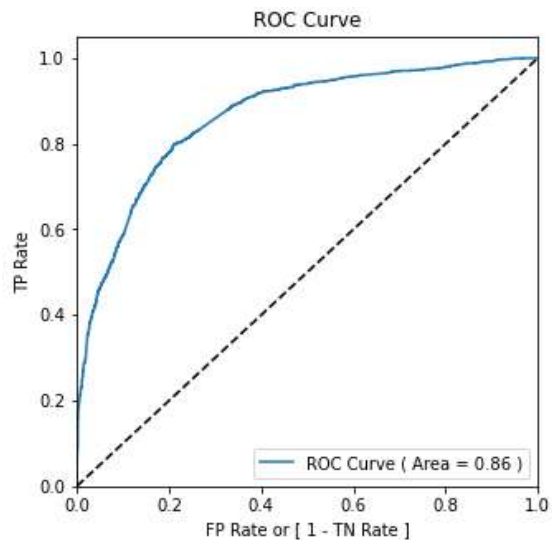
ROC curve after cut off optimization :
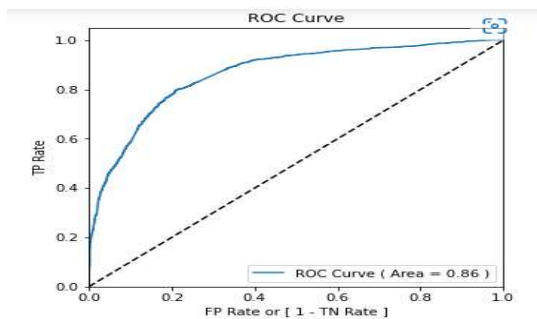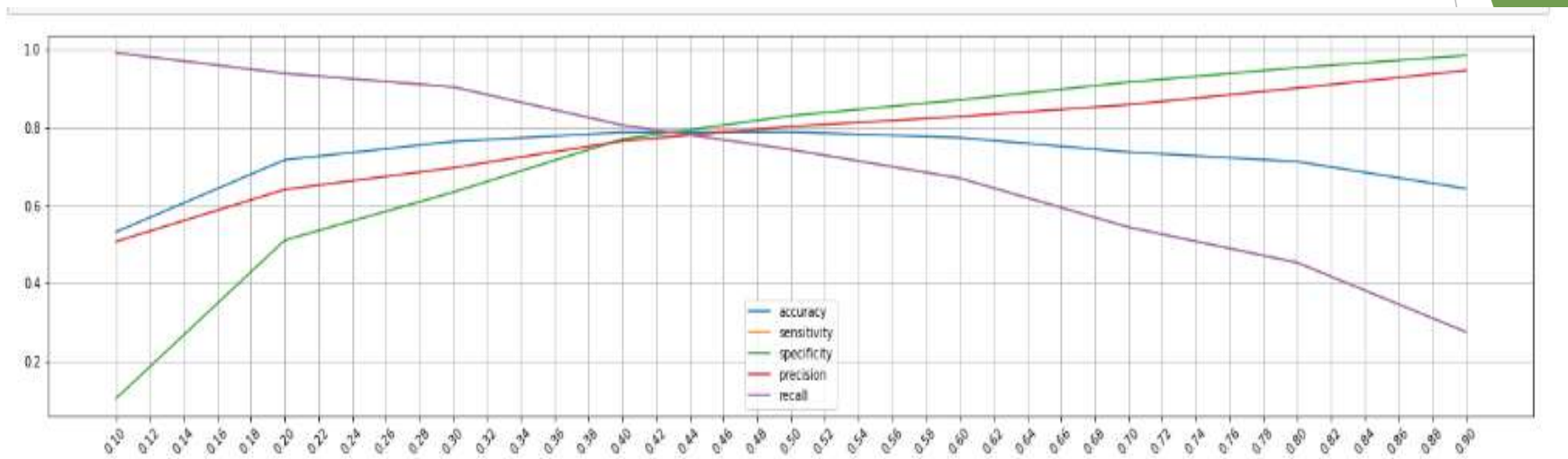
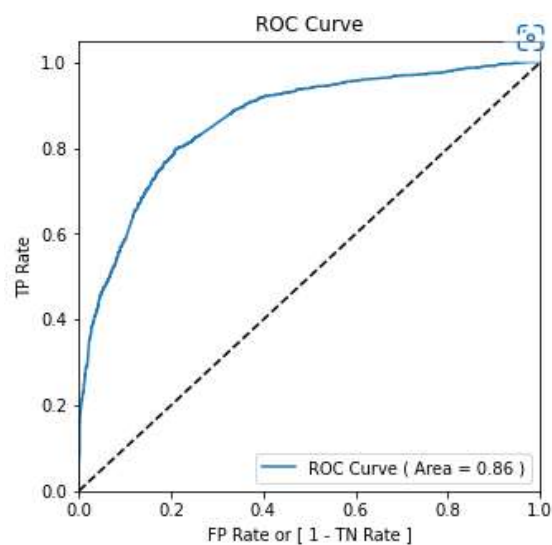As it is visible from above graph that optimal cut off is 0.43.

Confusion matrix

	Not Churn	Churn
Not Churn	1832	480
Churn	449	1700

	metric	train
0	accuracy	0.79
1	sensitivity	0.79
2	specificity	0.79
3	precision	0.78
4	recall	0.79

Model Evaluation – Sensitivity and Specificity on Test Dataset:

▶ ROC curve for test set:

Confusion Matrix

	Not Churn	Churn
Not Churn	785	211
Churn	198	718

	metric	test
0	accuracy	0.79
1	sensitivity	0.78
2	specificity	0.79
3	precision	0.77
4	recall	0.78

After comparing all metrics on train and test:

	metric	train	test
0	accuracy	0.79	0.79
1	sensitivity	0.79	0.78
2	specificity	0.79	0.79
3	precision	0.78	0.77
4	recall	0.79	0.78

Conclusion

Good Performance : The Model perform almost similar in the Train & Test Sets - Accuracy around 80%

Low Over fitting Tendency : This model does not seem to over fit to the training dataset

Impression : Hence Capable Making descent Prediction - Would Give confidence to the state-holders