LEAD SCORE CASE STUDY

Submitted by :

- Rashmi Singh
- Ratul Babbar
- Sheetal Ds Rathod

LEAD SCORE CASE STUDY FOR X EDUCATION

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

- Read the Data from Source
- Convert data into clean format suitable for analysis
- Remove duplicate data
- Outlier Treatment
- Exploratory Data Analysis
- Feature Standardization.

Feature Scaling and Splitting Train and Test Sets

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

Model Building

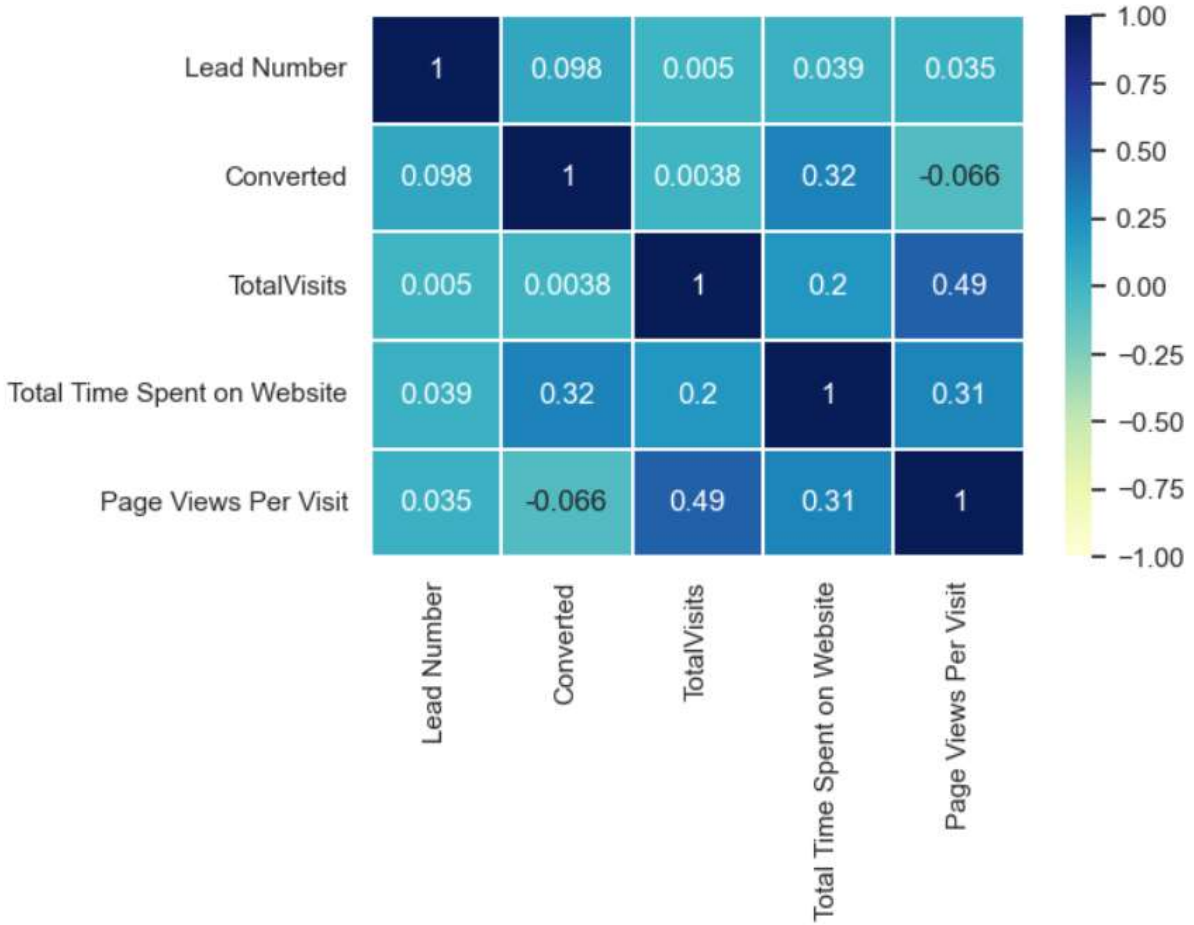
- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

Result

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

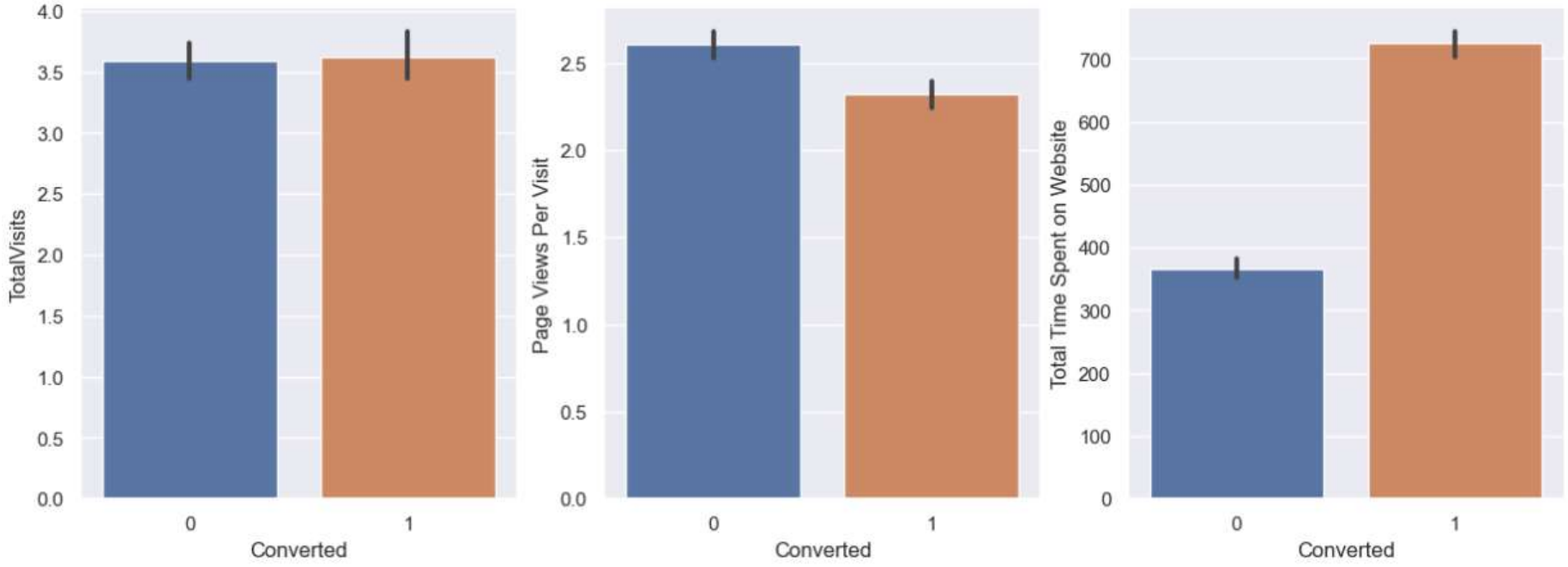
EDA

There is a strong positive correlation between **'Total Visits'** and **'Page Views per Visit'**, indicating customers visiting website tend to view more pages per visit



EDA

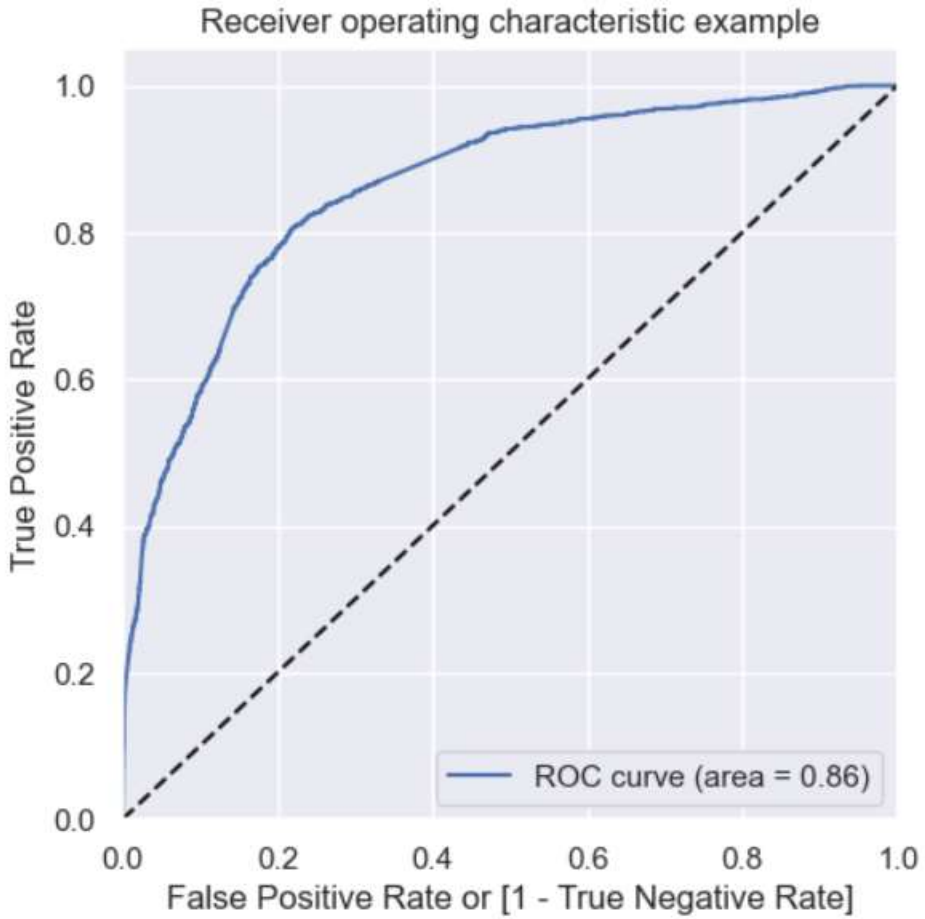
- Customers who spend more time on the website indicates increasing time spent on the website can lead to higher conversion rates.



VARIABLES IMPACTING THE CONVERSION RATE

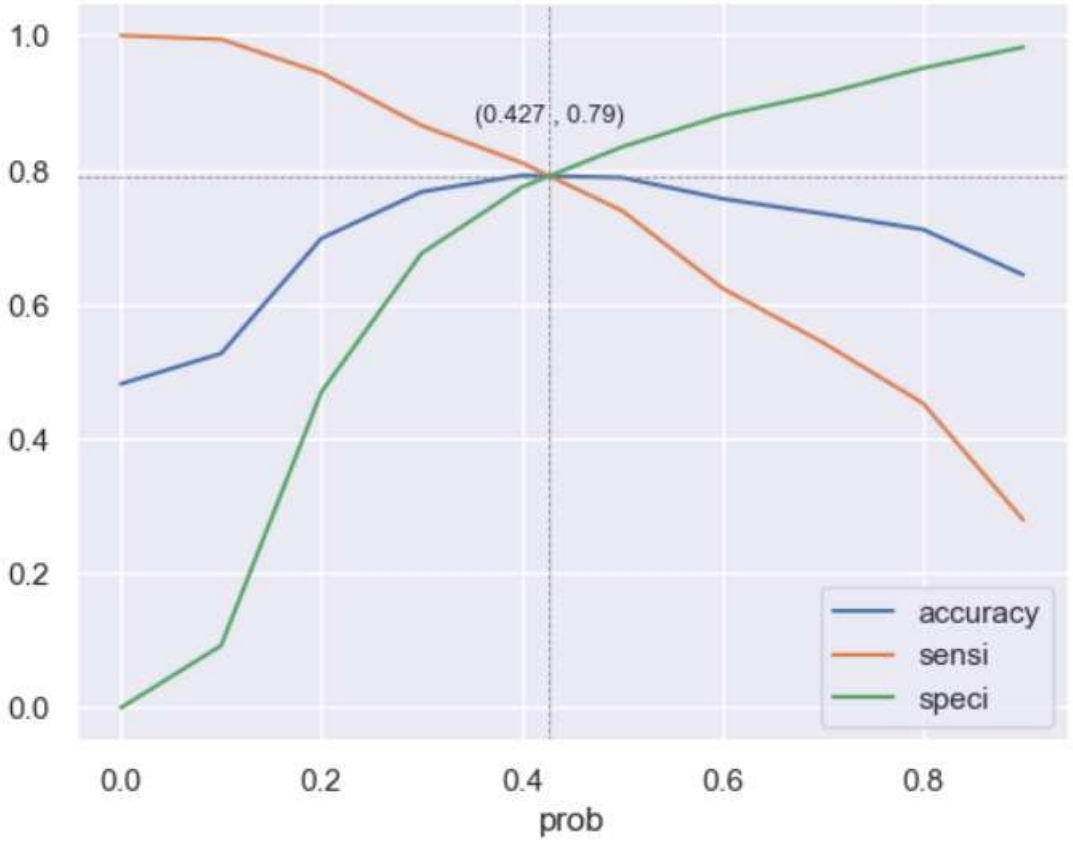
```
TotalVisits
Total Time Spent on Website
Lead Origin_Lead Add Form
Last Notable Activity_Unreachable
Last Activity_Had a Phone Conversation
Lead Source_Welingak Website
Lead Source_Olark Chat
Last Activity_SMS Sent
Do Not Email_Yes
What is your current occupation_Student
What is your current occupation Unemployed
```

MODEL EVALUATION – ROC (RECEIVER OPERATING CHARACTERISTIC)



The area under the curve of the ROC is 0.86

MODEL EVALUATION – SENSITIVITY AND SPECIFICITY ON TRAIN DATA SET



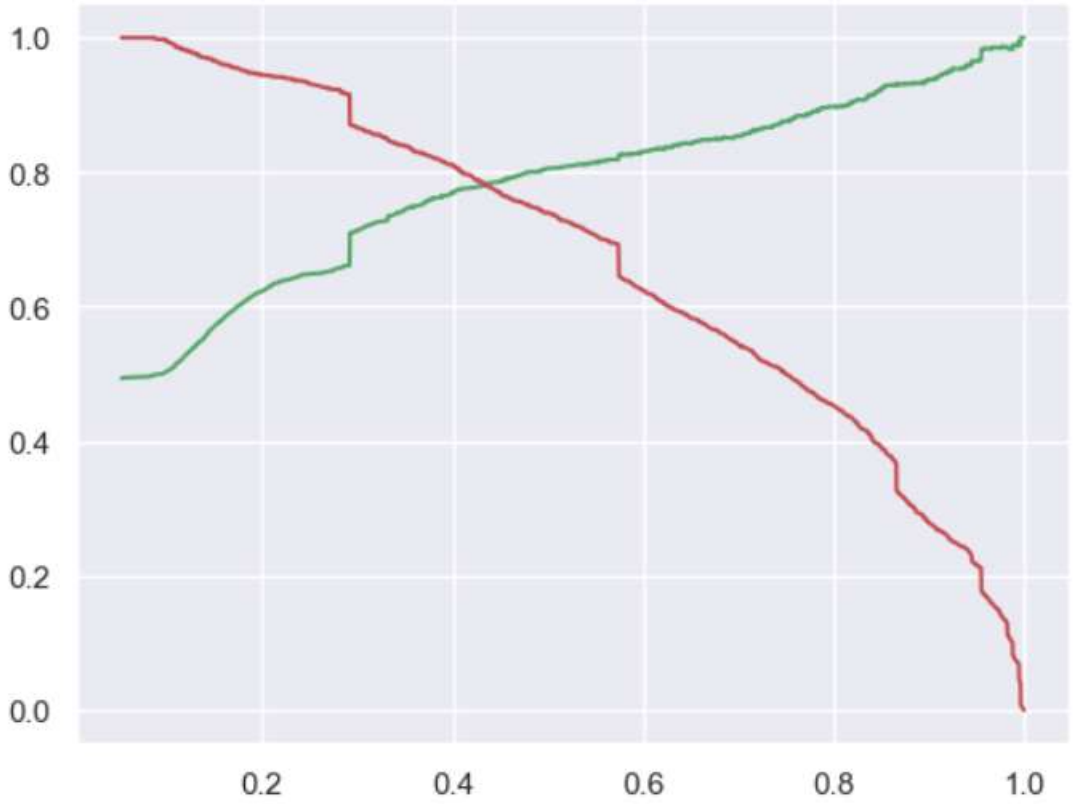
Confusion Matrix for Training data

1823	489
444	1705

- Accuracy – 79.08%
- Sensitivity – 79.34%
- Specificity – 78.84 %

Note : As the plot shows, we get the optimal values of the three metrics at around0.42, considering this as our cutoff now.

MODEL EVALUATION- PRECISION AND RECALL ON TRAIN DATASET



- Precision – 80.57 %
- Recall – 73.94 %

MODEL EVALUATION – CONCLUSION DRAWN FROM TRAIN-TEST DATASET

Train Data Set:
Accuracy: 79.08%
Precision: 77.71%
Recall: 79.33%

Test Data Set:
Accuracy: 63.02%
Precision: 59.35%
Recall: 67.65%

CONCLUSIVE OBSERVATIONS

Following are the important variables contributing towards the probability of lead conversion.

TotalVisits 11.148912

Total Time Spent on Website 4.422291

Lead Origin_Lead Add Form 4.205123

Lead Origin_Lead Add Form, Lead Source_Olark Chat and Last Activity_Had a Phone Conversation are the categorical / dummy variables in the model to be focused to improve the probability of lead conversion.

0.38 is the tradeoff between Precision and Recall - Thus we can safely choose to consider any Prospect Lead with Conversion Probability higher than 38%

List of features to be considered : Lead Source_Olark Chat, Specialization_Others, Lead Origin_Lead Add Form, Lead Source_Welingak Website, Total Time Spent on Website, Lead Origin_Landing Page Submission, What is your current occupation_Working Professionals, Do Not Email

The conversion rate is 30-35% (close to average) for API and Landing page submission More number of leads are generated by google / direct traffic. Maximum conversion ratio is by reference and welingak website