upGrad

Lead Score Case Study Using Regression Model

Submitted by :

Deepak Palsavdiya

Deepika M

Lead Score Case Study forX 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 - Accuracy, 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.
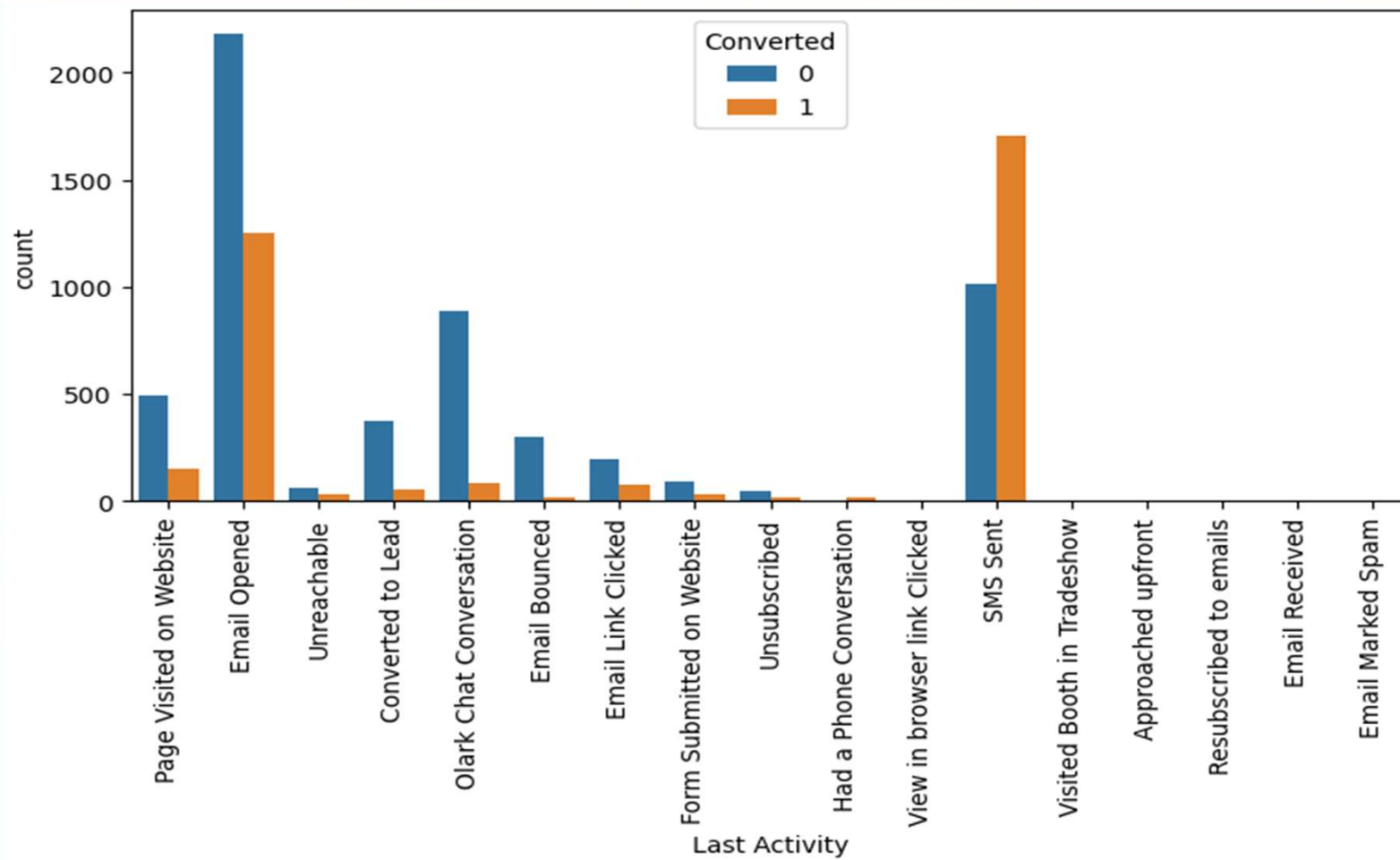
Result

- 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

In Lead Origin, maximum conversion happened from Landing Page
Submission
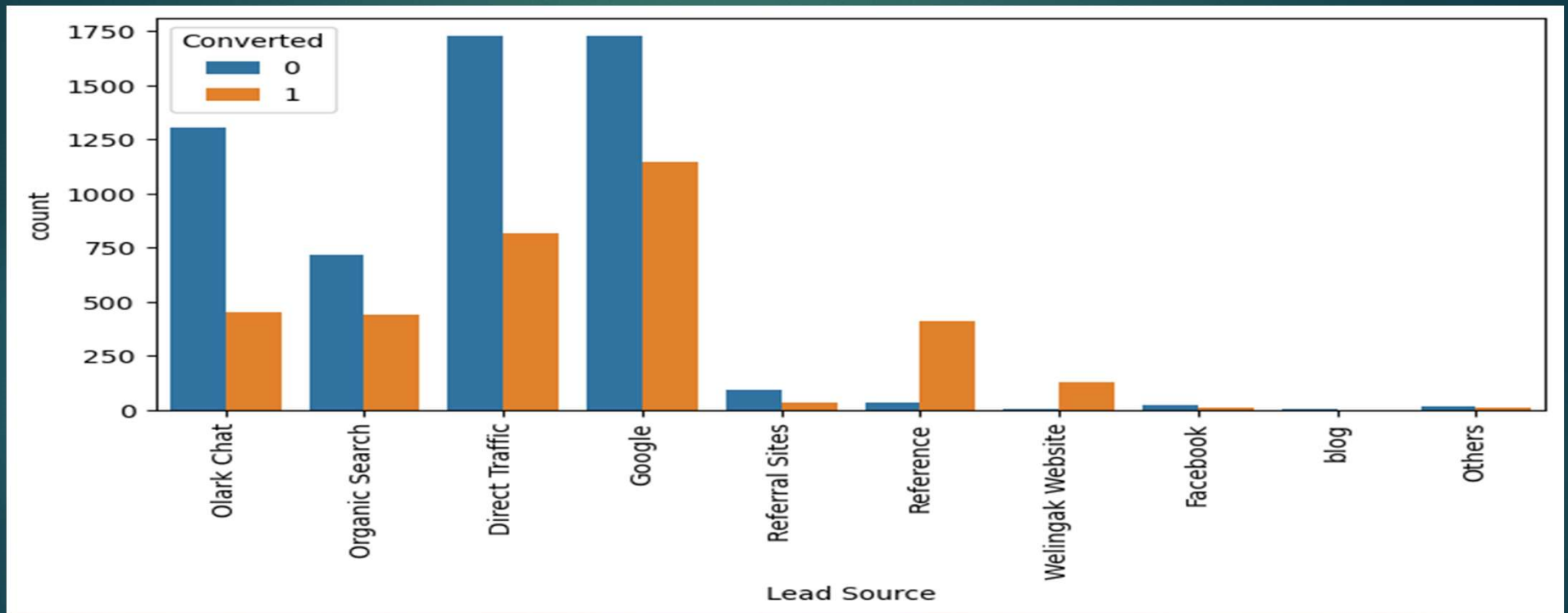
Major conversion in the lead source is from Google

Last Activity value of SMS Sent' had more conversion.

Variables Impacting the Conversion Rate

- Lead Origin_Landing Page Submission
- Specialization_others
- Last Activity_Email Opened
- Last Activity_SMS Sent
- Lead Source_Olark Chat
- Last Activity_Olark Chat Conversation
- Lead Origin_Lead Add Form
- Lead Source_Welingak Website
- Total Time Spent on Website
- Do Not Email
- What is your current occupation_Working Professional
- Last Activity_Others

Model Evaluation - Sensitivity and Specificity on Train Data Set

The graph depicts an optimal cut off of 0.35 based on Accuracy,
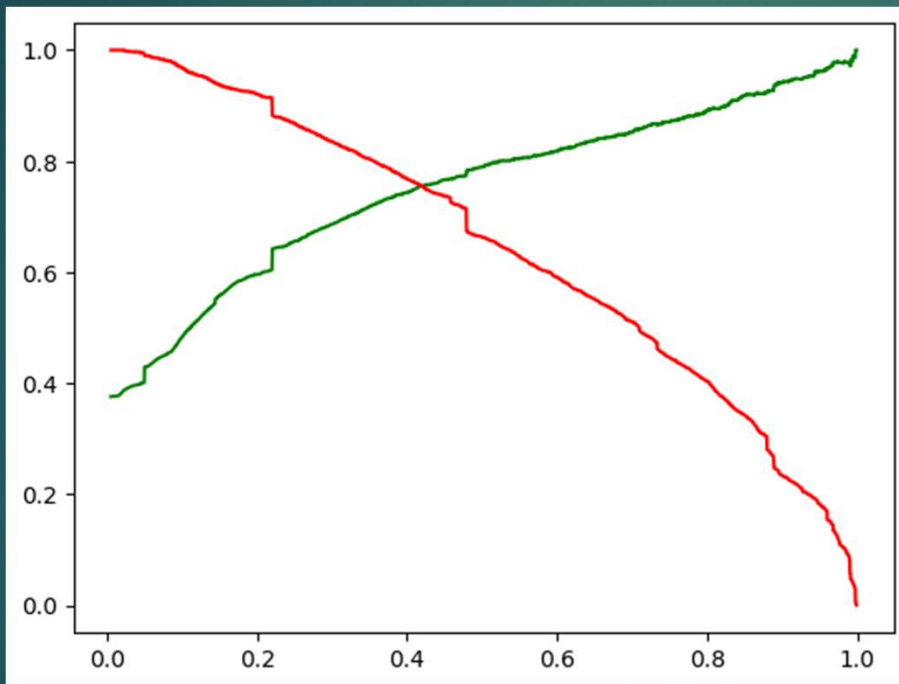Sensitivity and Specificity
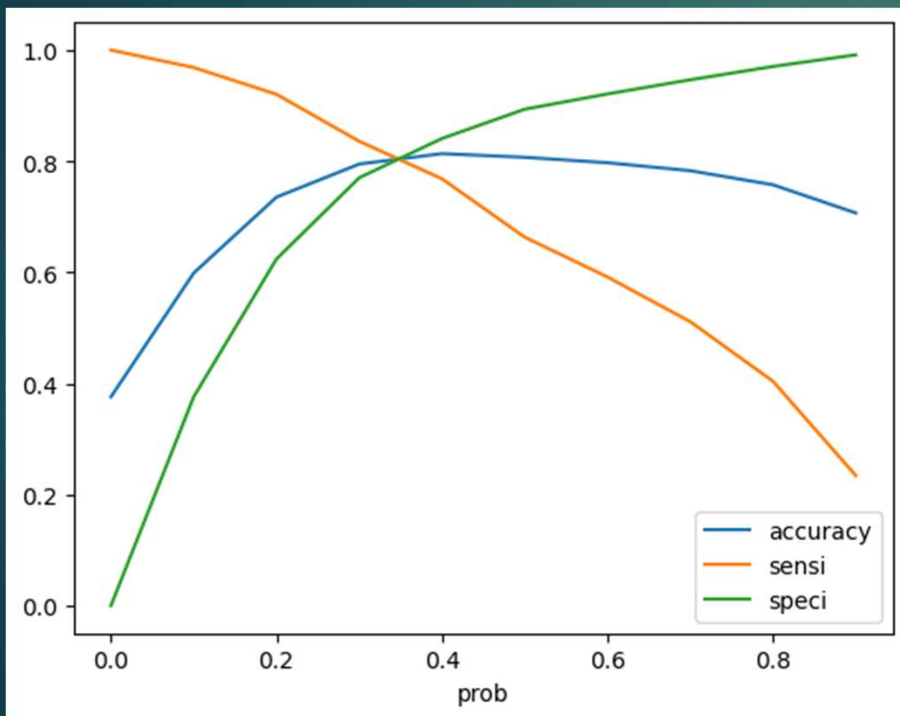
- Accuracy - 80%
- Sensitivity - 80 %
- Specificity - 81 %
- False Positive Rate - 18 %
- Positive Predictive Value - 71 %
- Positive Predictive Value – 87%

Model Evaluation- Precision and Recall on Train Dataset

The graph depicts an optimal cut off of 0.42 based on Precision and Recall

- Precision - 78.90 %
- Recall - 66.31%

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. –

	Train data.	Test Data
Accuracy :	80%	81%
Sensitivity :	80%	80%
Specificity :	81%	81%

➢ 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

➢ The top 3 variables that contribute for lead getting converted in the model are
  ➢ Total time spent on website
  ➢ Lead Add Form from Lead Origin
  ➢ Had a Phone Conversation from Last Notable Activity

➢ Hence overall this model seems to be good.

THANKS