LEAD SCORING CASE STUDY

Submitted by:
Trupti Sudhir

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.

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 lead, i.e., the leads that are most likely to convert to a paying customer.

The company requires us to build a model wherein we 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 has given a ballpark of the target lead conversion rate to be around 80%.

Overall Strategy:

- Source the data for analysis
- Clean and prepare the data
- Exploratory Data Analysis
- Feature scaling
- Splitting the data into test and train datasets
- Building a logistic regression model and calculate the ‘Lead score’
- Evaluating the model by using different metrics – Sensitivity, specificity, precision and recall
- Applying the best model in test data based on sensitivity and specificity metrics

Problem Solving Methodology

Data Sourcing, cleaning and prep

- Read the data from source .csv file
- Convert data into format suitable for analysis
- Remove duplicate data
- Outlier treatment of data
- Exploratory Data Analysis
- Feature standardization

Feature Scaling, Splitting Train and test datasets

- Feature scaling of numeric data
- Splitting data into train and test set.

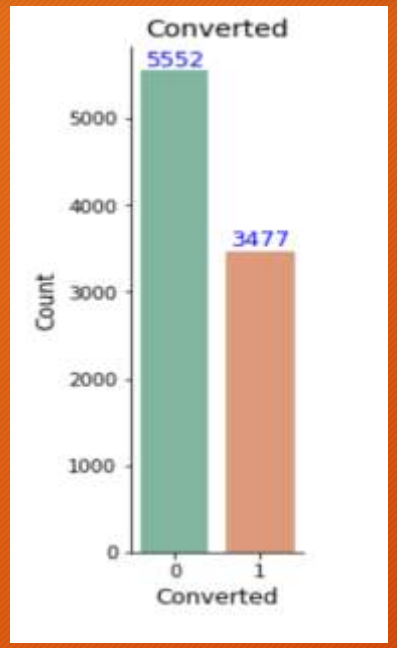
Model Building

- Feature section using RFE
- Determine optimal model using logistic regression
- Calculate metrics like accuracy, sensitivity, specificity, precision and recall. Evaluate the model using those.

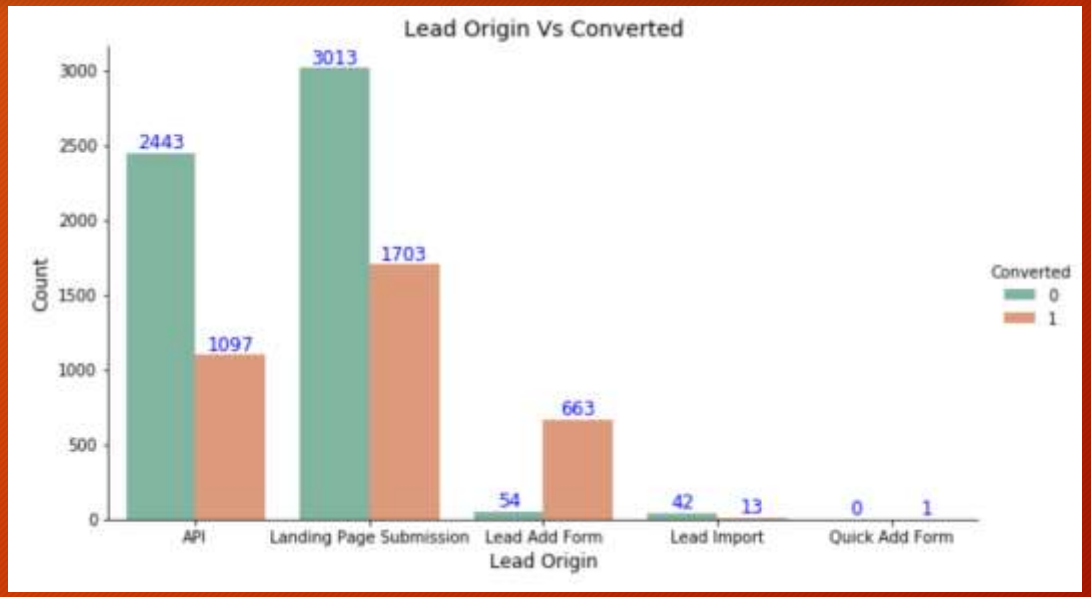
Result

- Determine lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cutoff threshold.

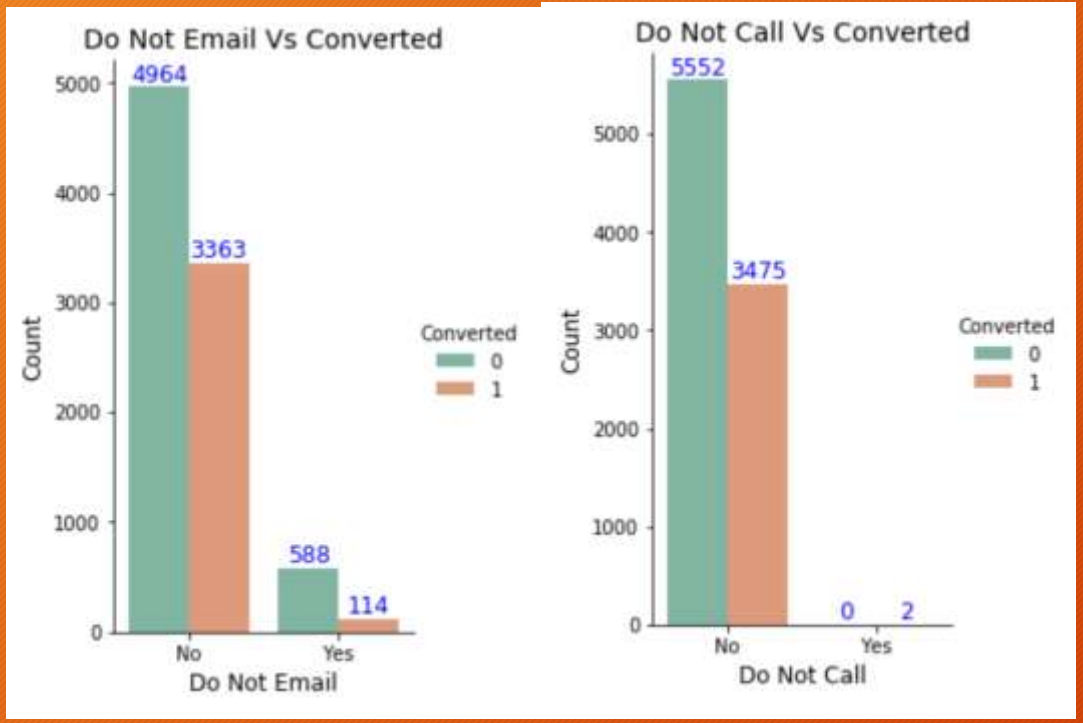
EDA:



We have around 39% (3477/9092) conversion rate in total.

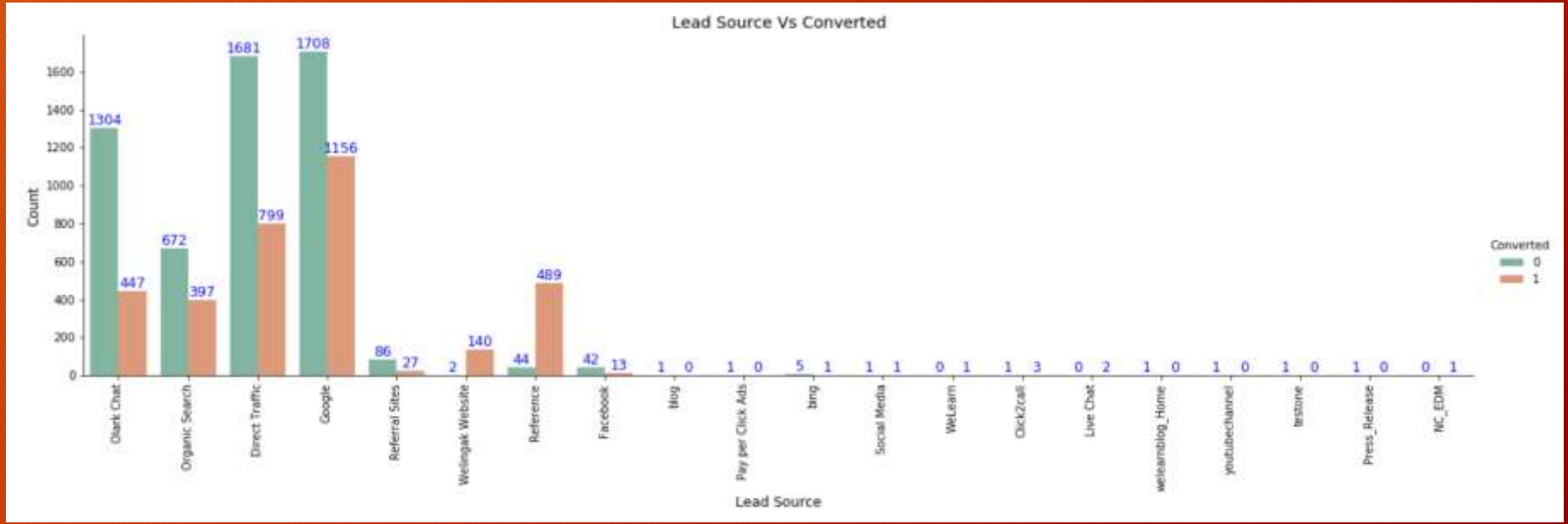


In Lead origin, maximum conversion is from Landing Page Submission.



Major conversion is from emails sent and calls made.

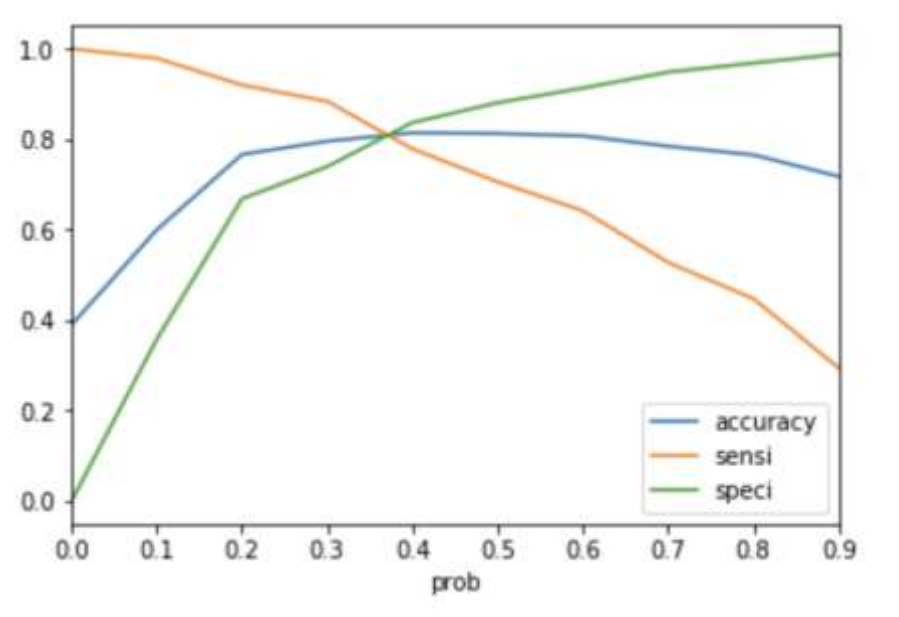
It can be observed that major conversion in the lead source is from google (1156)



Variables impacting the Conversion Rate:

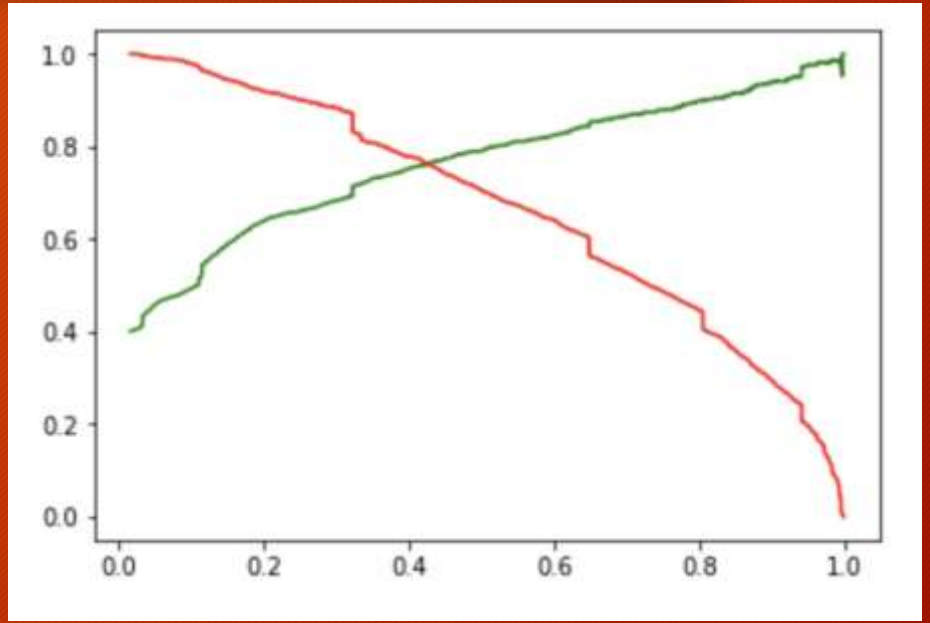
- Do not Email
- Total Visits
- Total time spent on website
- Lead Origin – Lead Page Submission
- Lead Origin – Lead Add Form
- Lead Source - Olark Chat
- Last Source – Welingak Website
- Last Activity – Email Bounced
- Last Activity – Not Sure
- Last Activity – Olark Chat Conversation
- Last Activity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

Model Evaluation (Train Data set):



The graph depicts an optimal cutoff of 0.37 based on accuracy, sensitivity and specificity.

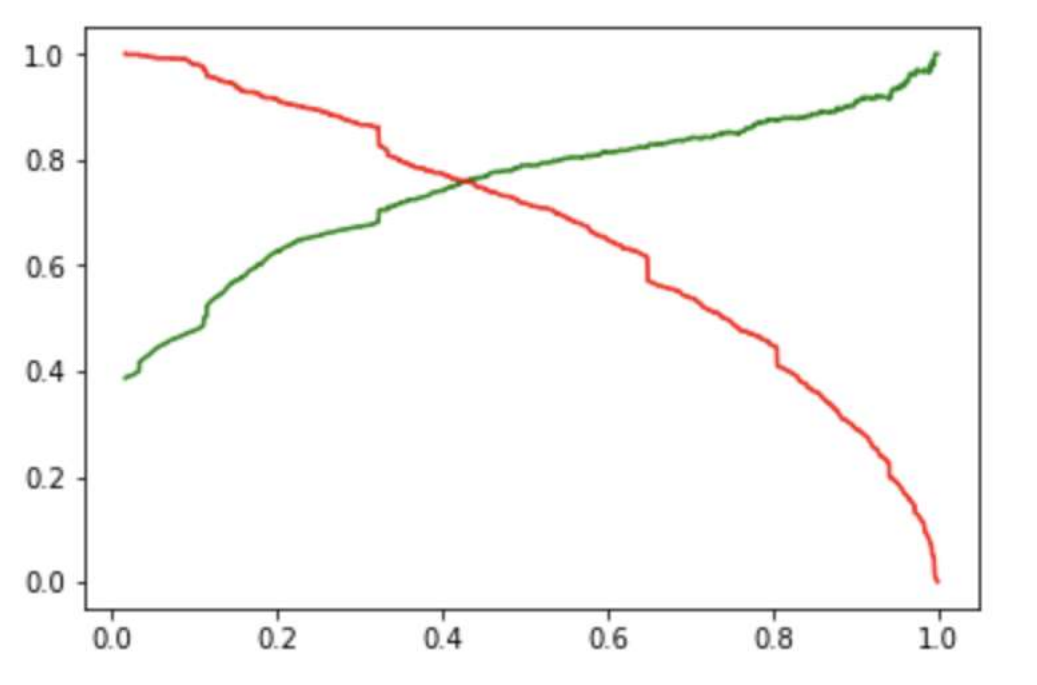
Accuracy – 81%
Sensitivity – 80%
Specificity – 82%
False positive rate – 18%
Positive Pred vale – 74%



The graph depicts an optimal cutoff of 0.42 based on Precision and recall.

Precision – 79%
Recall – 71%

Model Evaluation (Test Data set):



Accuracy – 80%
Sensitivity – 79%
Specificity – 82%
Precision – 73%
Recall – 79%

The graph depicts an optimal cutoff of 0.42 based on Precision and recall.

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 81%, 79% and 82% which are approximately closer to the respective values calculated using trained set.

The lead score calculated shows the conversion rate on the final predicted model is around 80% (in train set) and 81% 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.