Lead Score Case Study

Submitted By :
Jagruti Desai
Saeed Ahmad Khan
Jyoti Bhatnagar

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%.

Goal for this case study : Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Strategy

- Import Data
- Clean and prepare the acquired data for further analysis
- Exploratory data analysis for figuring out most helpful attributes for conversion
- Scaling features
- Prepare the data for model building
- Build a logistics regression model
- Assign a lead score for each leads
- Test the model on train set
- Evaluate the model by different measures and metrics
- Test the model on test set
- Measure the accuracy of the model and other metrics for evaluation

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
- EDA
- Feature standardization

Feature Scaling and splitting train and test

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

Model Building

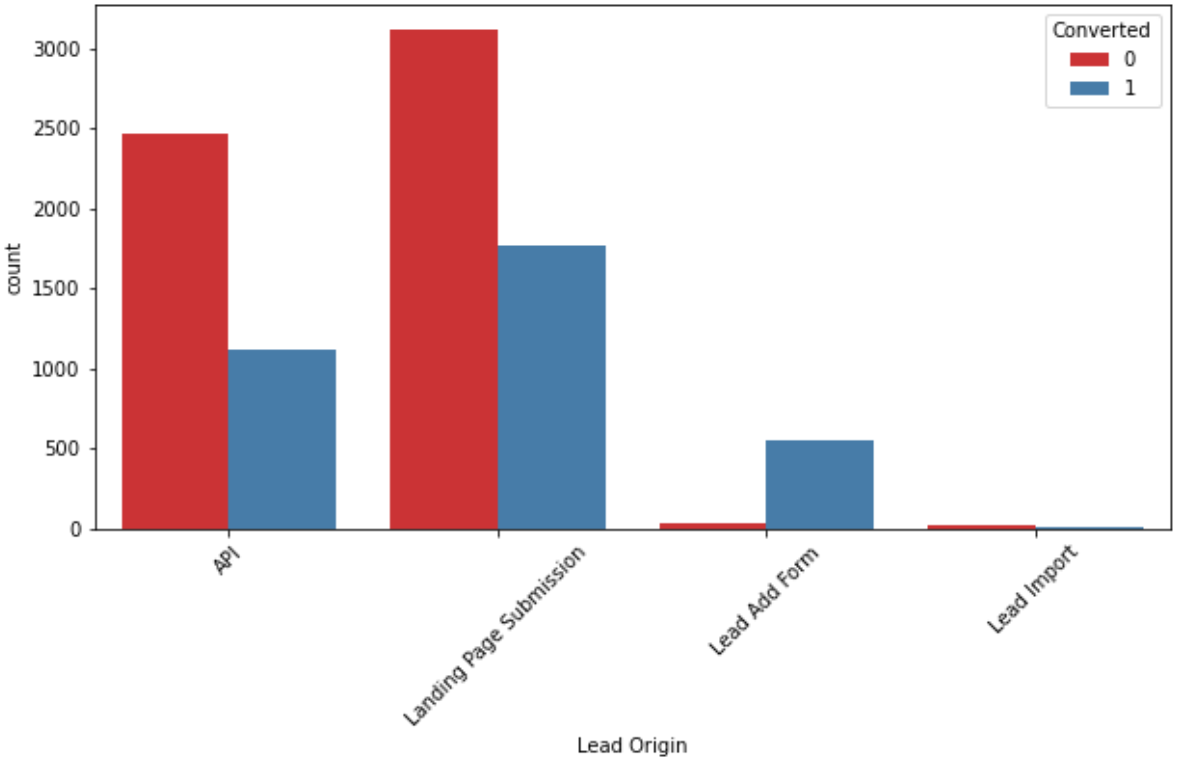
- Feature selection using RFE
- Determine the optimal model using Logistics regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model

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

Exploratory Data Analysis

Lead conversion rate is 38%



LEAD ORIGIN

Inference :

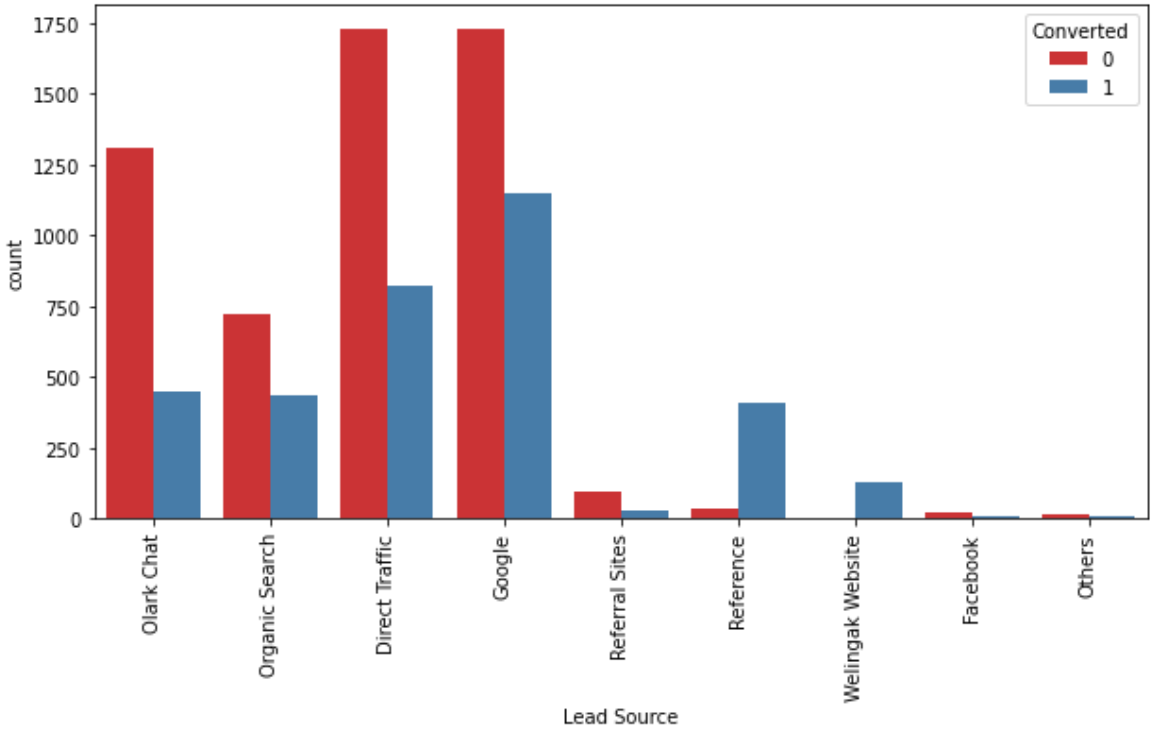
API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.

Lead Add Form has more than 90% conversion rate but count of lead are not very high.

Lead Import are very less in count.

To improve overall lead conversion rate, we need to focus more on improving lead converion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

Exploratory Data Analysis



LEAD SOURCE

Inference

Google and Direct traffic generates maximum number of leads.

Conversion Rate of reference leads and leads through welingak website is high.

To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

X EDUCATION FORUMS, NEWSPAPERS, DIGITAL ADVERTISEMENT, THROUGH RECOMMENADTIONS, RECEIVE MORE UPDATES ABOUT OUR COURSES, MAGAZINES, NEWSPAPER ARETICLES and SEARCH

Inference :

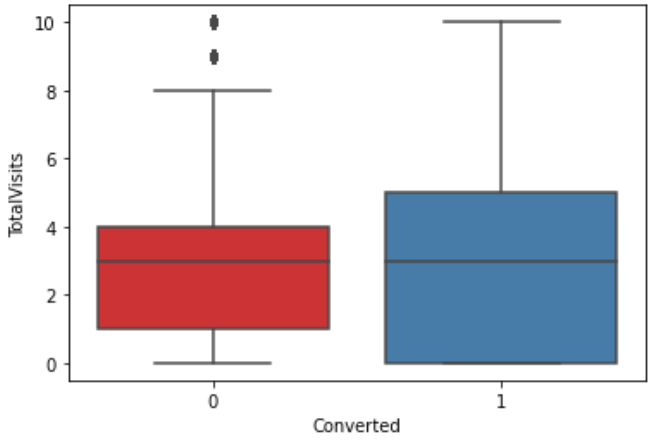
Most entries above are 'No'. No Inference can be drawn with these parameter.

Exploratory Data Analysis

Total Visits

Median for converted and not converted leads are the same.

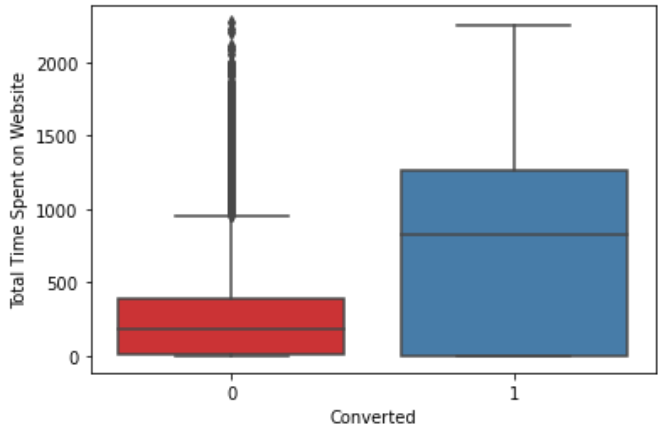
Nothing can be concluded on the basis of Total Visits



Total time spent on Website

Leads spending more time on the weblise are more likely to be converted.

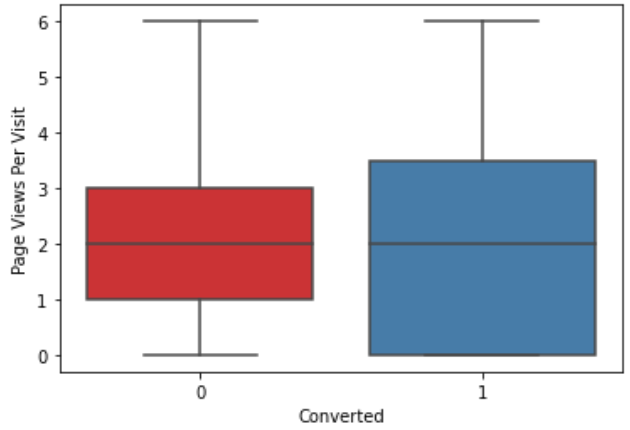
Website should be made more engaging to make leads spend more time



Page views per visit

Median for converted and unconverted leads is the same.

Nothing can be said specifically for lead conversion from Page Views Per Visit

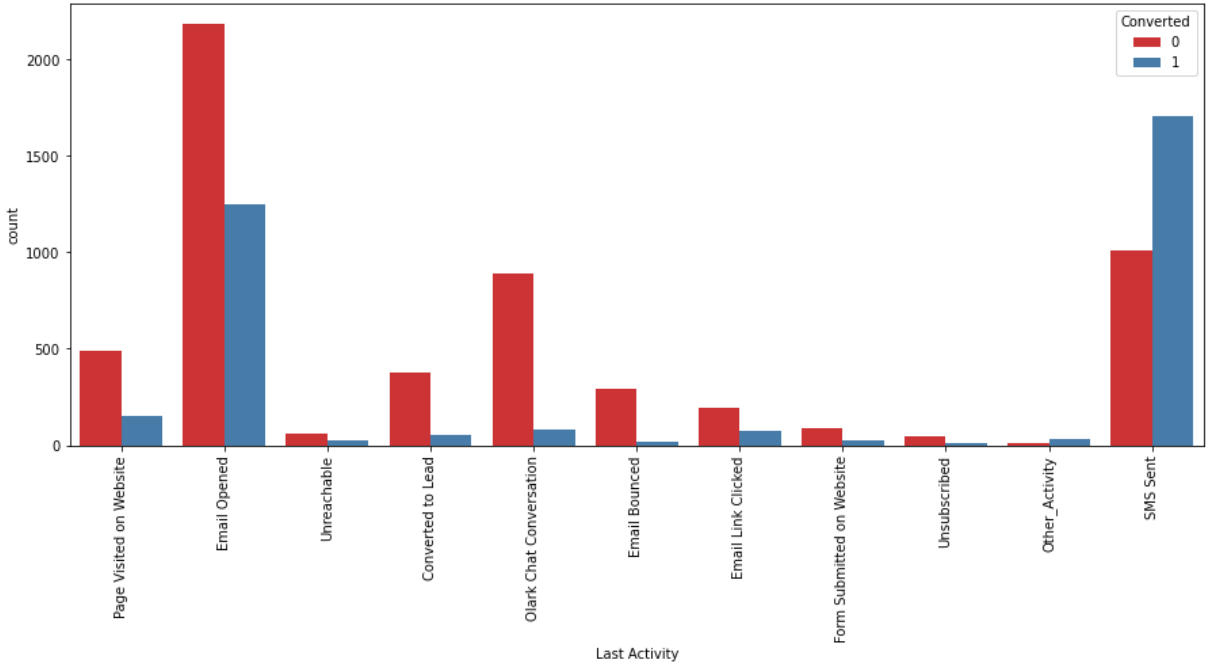


Exploratory Data Analysis

Last Activity

Most of the lead have their Email opened as their last activity.

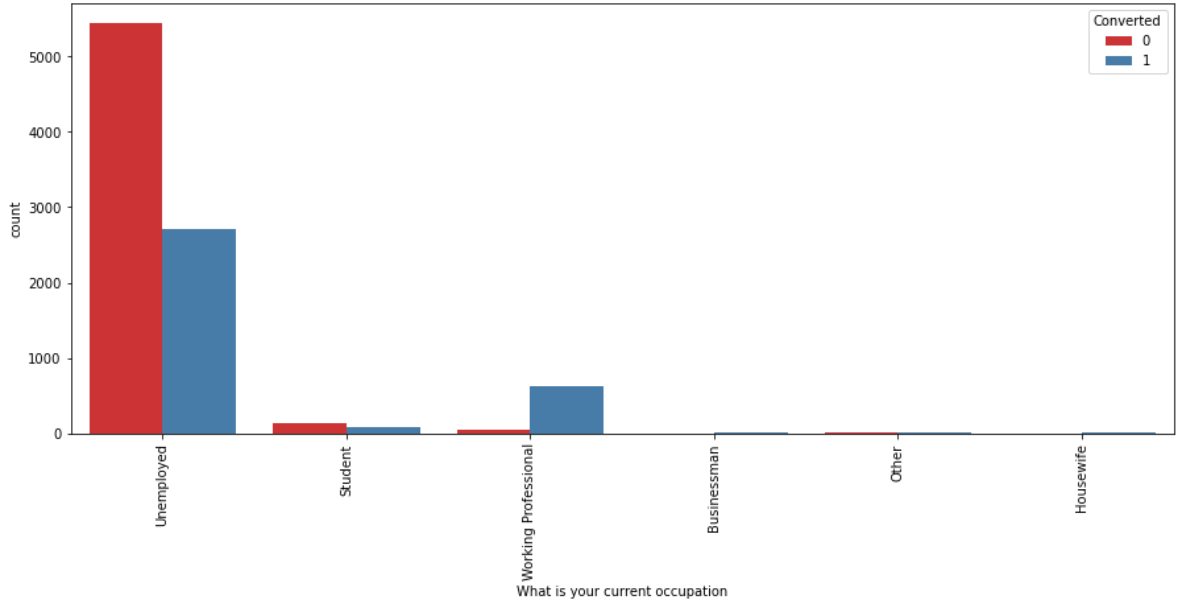
Conversion rate for leads with last activity as SMS Sent is almost 60%.



What is your current occupation

Working Professionals going for the course have high chances of joining it.

Unemployed leads are the most in numbers but has around 30-35% conversion rate.

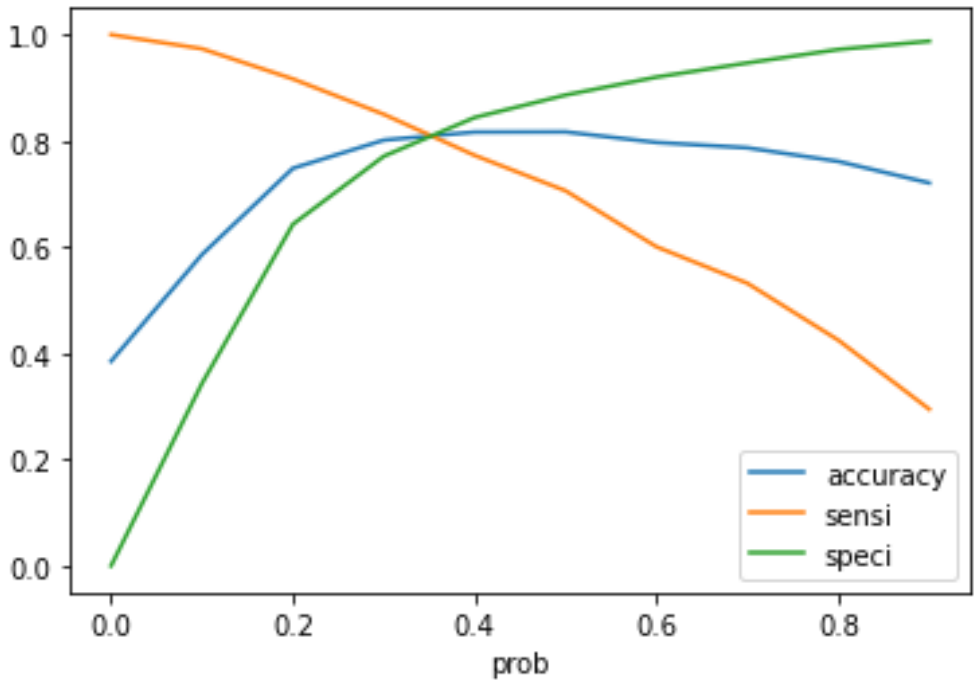


Model building

- Splitting into train and test set
- Scale variables in train set
- Build the first model
- Use RFE to eliminate less relevant variables
- Build the next model
- Eliminate variable based on high p-values
- Check VIF value for all the exiting columns
- Predict using train set
- Evaluate accuracy and other metric
- Predict using test set
- Precision and recall analysis on test predictions

Model evaluation – sensitivity and specificity on train data set

The graph depicts an optimal cut off of 0.34 based on Accuracy, Sensitivity and Specificity

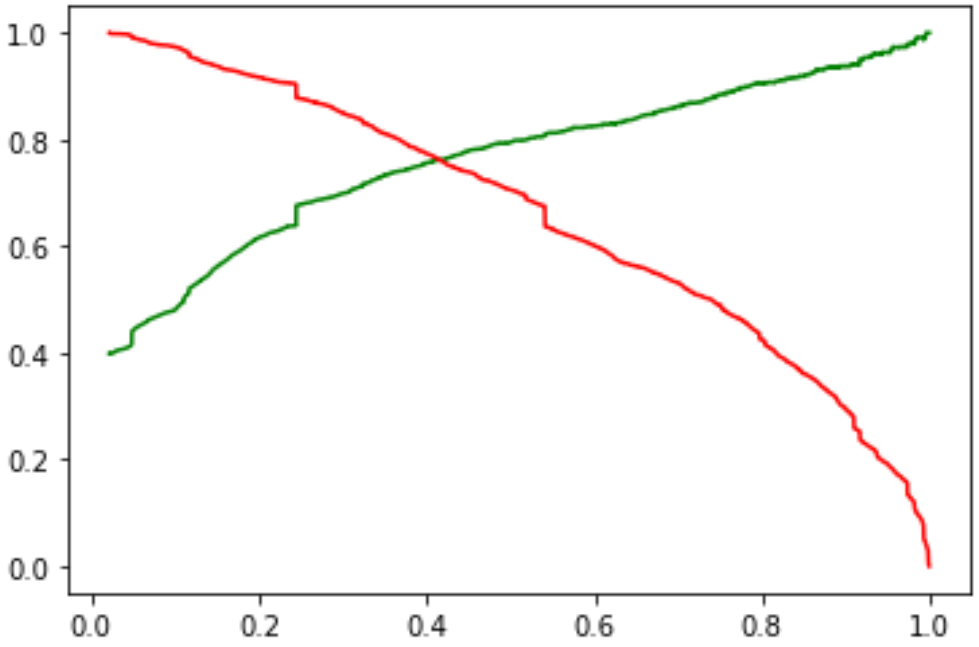


Confusion matrix
3151, 754
447, 1999

Accuracy : 81%
Sensitivity : 81.7%
Specificity : 80.6 %
False positive rate : 19.3%
Positive predictive value : 72.6%
negative predictive value : 87.5%

Model evaluation – precision and recall on train data set

The graph shows the trade-off between the Precision and Recall



Confusion matrix
3461, 444
719, 1727

Precision : 79.5%
Recall : 70.6%

Conclusion

Observation :

After running the model on the Test Data , we obtain:

Accuracy : 80.4 % Sensitivity : 80.4 % Specificity : 80.5 %

Results:

(1) Comparing the values obtained for Train & Test:

Train Data:

Accuracy : 81.0 % Sensitivity : 81.7 % Specificity : 80.6 %

Test Data:

Accuracy : 80.4 % Sensitivity : 80.4 % Specificity : 80.5 %

Thus we have achieved our goal of getting a ballpark of the target lead conversion rate to be around 80% . The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model to get a higher lead conversion rate of 80%

(2) Finding out the leads which should be contacted:

The customers which should be contacted are the customers whose "Lead Score" is equal to or greater than 85. They can be termed as 'Hot Leads'.