

# Lead Scoring Case Study Summary

By

Subhranil Sengupta and Steve Thomas

## Problem Statement:

X Education, an online education provider, faces low lead conversion challenges despite a steady influx of leads. To address this, the company aims to identify 'Hot Leads' – those with the highest potential for conversion. By implementing strategies such as lead scoring, data analysis, behavioral tracking, and personalized communication, X Education seeks to prioritize engagement with the most promising prospects. Additionally, optimizing the referral program and providing sales team training will enhance efficiency in lead management. Through continuous monitoring and refinement of these strategies, X Education aims to boost its lead conversion rate, drive revenue growth, and improve overall business performance.

## Steps Followed:

### 1. Data Understanding

### 2. Data Cleaning/Manipulation: Removed Columns having more than 45% Null values.

Here the following columns have more than 45% null values:

- (I.) Lead Quality
- (ii.) Asymmetrique Activity Index
- (iii.) Asymmetrique Profile Index
- (iv.) Asymmetrique Activity Score
- (v.) Asymmetrique Profile Score

These are the columns we are going to drop.

### 3. Exploratory Data Analysis:

(i.) Data Imbalance Checking on 'Converted' Variable: Successful lead conversion rate is just 42.3%. But, 57.7% of the Leads have not converted. So, the data is imbalanced.

(ii.) Univariate Analysis: Observed the following from Univariate Analysis

(a.) Do not call: 99.97% of all the leads do not want to get a call.

(b.) Do not email: 7.34% of all the leads prefer to get an email.

(c.) Lead Source: Most leads come from Google, Direct Traffic and Organic Search. X Education should put more emphasis on the mentioned sources while trying to increase the number of lists coming from Ads, Social Media, and Press Release.

(d.) Country: 95% of the leads come from India.

(e.) Specialization: Here the selected specialization seems distributed between Finance Management, Human Resource Management, Marketing Management, Operations Management, Business Administration, IT Projects Management, Supply Chain Management, Banking, Investment And Insurance, Media and Advertising. However, it looks like almost 15% of people did not select any specialization.

(f.) How did you hear about X: In the case of the column "How did you hear about X Education" as well we see that most leads have not selected a valid option.

**4. Bivariate Analysis:** Observed the following from Bivariate Analysis

- (i.) Lead Source: Most leads come from Organic Search, Direct Traffic and Google. It is also important to notice that Reference has a very high conversion rate.
- (ii.) What is your current occupation: From the plot, we can see that most leads are Unemployed. However, Working Professionals have the highest conversion rate.
- (iii.) Last Activity: "SMS Sent" has the highest lead conversion rate.
- (iv.) Lead Origin: Most leads come from Landing Page Submission. But Lead Add Form has a higher lead conversion rate than any other category.

**5. Data Transformation:** Changed Multi-category variables into dummies and binary variables('Yes'/'No') to ('1'/'0')

**6. Data Preparation:**

- (i.) Split the data into Train and Test Datasets.
- (ii.) Performed Feature Scaling

**7. Model Building:**

- (I.) Used RFE for Feature Selection
- (II.) Selected 25 features using RF

**8. Evaluation Metrics are:**

**(i.) Train Data:**

- (a.) Model Accuracy: 96.21 %
- (b.) Model Sensitivity: 96.69 %
- (c.) Model Specificity: 95.86 %
- (d.) Model Precision: 94.43 %
- (e.) Model Recall: 96.69 %
- (f.) Model F1-Score: 95.55 %

**(ii.) Test Data:**

- (a.) Model Accuracy value: 94.56 %
- (b.) Model Sensitivity value: 95.71%
- (c.) Model Specificity value: 93.69 %
- (d.) Model Precision: 91.94 %
- (e.) Model Recall: 95.71 %
- (f.) Model F1-Score: 93.79 %

Hence, our model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in reaching his goal to achieve 80% lead conversions with this model.

**9. Model Parameters:**

**(i.) The top 3 positively contributing features that predict hot leads in the model are:**

- (a.) Tags\_Closed by Horizzon
- (b.) Tags\_Will revert after reading the email
- (c.) Country\_Hong Kong

(ii.) **The Top 3 negatively contributing features that predict hot leads in the model are:**

- (a.) Tags\_switched off
- (b.) Tags\_Not doing further education
- (c.) Tags\_invalid number

## **Recommendations:**

### **1. From Exploratory Data Analysis:**

(i.) When the lead source is Reference, the conversion rate is high. X Education should focus on focus on References. Providing a referral bonus can be beneficial.

(ii.) Working professionals are more likely to convert than Unemployed People and students. The marketing team should make campaigns targeting working professionals.

(iii.) Leads that received SMS have the highest lead conversion rate. X Education should focus on reaching out to the leads via messaging services.

### **2. From Logistic Regression Model:**

(ii.) Other than India, focus on Hong Kong, and Germany for potential leads.

(iii.) People who spend more time on the website are more likely to be a 'Hot Lead'. Making the website more engaging or bug-free is important.

(iv.) People who are interested in MBA or other courses or not interested in higher education, will most likely won't be converted.

(v.) School students have less provability to take any course from X Education. Maybe all the courses are designed for graduate college students or working professionals.

## **Conclusion:**

The evaluation metrics are pretty close to each other so it indicates that the model is performing consistently across different evaluation metrics in both test and train datasets.

- The model achieved a sensitivity of 96.22% in the train set and 94.56% in the test set, using a cut-off value of 0.3
- Sensitivity in this case indicates how many leads the model identifies correctly out of all potential leads that are converting
- The CEO of X Education had set a target sensitivity of around 80%
- The model also achieved an accuracy of 94.56% and a Recall of 95.72%, which is in line with the study's objectives