# LEAD SCORING CASE STUDY

# PROBLEM STATEMENT

An education company named X sells online courses to industry professionals on several websites and search engines like Google.

Although X Education gets a lot of leads (When a person fill up a

form providing his/her email address or phone number,

he/she is classified to be a lead), its lead conversion rate is very poor

(as shown in the attached image).

To make the lead selection process more efficient, the company wishes to identify the most potential leads- ‘Hot Leads’.

This will help in several ways

1. Time and cost saving
2. Less resource required.
3. Focus on other objective(sales team can now focus on communicating with the potential leads rather than making calls).

# APPROACH

To identify the ‘hot leads’ successfully, a well defined Logistic Regression model should be prepared.

Below are the files which were provided for analysis :

1. Leads.csv – It consists of almost 9k users data for analysis purpose

2. Leads Data Dictionary.xlsx – This file is the data dictionary of Leads.csv i.e. meaning of all the fields

# ANALYSIS

Following are the steps followed to prepare a goof-fit Logistic Regression model for Lead selection :

Step 1 : Reading and understanding the data

Step 2 : Preparing the data - below are some of the sub-steps

a. Converting binary variables (Yes/No) to 0/1

b. creating dummy variables for categorical values

c. handling null values

etc.

Step 3: Test-Train Split - Split the records into train and test data set.

Train Data Set - used for training the model

Test Data Set - used to test the prepared model and check it's accuracy and other metric values

Step 4: Feature Scaling and feature selection using RFE

Step 5: Start prediction model on Train Data Set

a. Initially, check with cut-off=0.5 and validate using different metrics (accuracy, precision, recall etc.)

b. Plot accuracy, sensitivity and specificity curve to get better-cut-off point, in this case measured cut-off was 0.38

c. Again run the predictions and check the metric values

Step 6: Predict on Test Data Set

# RESULTS

Below are some of the metric values calculated on Test Data Set :

Accuracy = 81%

Sensitivity = 79%

Specificity = 82%

Precision = 74%

Recall = 79%

It shows that the model is predicting potential Leads Selection efficiently.

# CONCLUSION

1. Working professional are more likely to enroll for courses.
2. People who visit the sites more often and spend more time searching different courses may want to enroll more.
3. Leads who have filled App Form are more likely to enrol.