# SUMMARY OF LEAD SCORING CASE STUDY

# X EDUCATION

Building a Logistic Regression Model to filter out the HOT Leads to focus more on them and thus enhancing the Conversion Ratio for X Education Company.

## 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. 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%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone .There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc. ) in order to get a higher lead conversion. X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you 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, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

# Steps Followed

## 1.Data loading and understanding

Total 9240 rows and 37 columns consisting 7 Numeric and 30 Categorical columns.

- we have to modify the column names
- Prospect ID and Lead Number are unique identifiers. They both have the same purpose hence it's better to drop one of them (Prospect ID)
- few columns have "select" as their entries. These are essentially null values because select appears when someone does not select anything from the dropdown

## 2.Data Cleaning

- Renaming the Columns
- Dropping the prospect_id column
- Replacing 'select' category with null values
- Handling the Null values

## 3.Exploratory Data Analysis

- Univariate Analysis on Numeric columns
- Univariate Analysis on Categorical Columns
- Bivariate Analysis on Numerical Columns
- Bivariate Analysis on Categorical Columns
- Find out presence of outliers
- Heat map
    - There is a much lower positive correlation between total_visits and converted
    - There is positive correlation between Total Time Spent on Website and converted
    - There is low negative correlation between Page Views Per Visit and Total Visits with Conversion

## 4.Data Preparation

- Converting Binary columns (Yes/No) to 0/1
- Creating dummy variable for categorical columns
- Test-Train Split
- Feature Scaling
- Drop highly correlated dummy variables

## 5.Model Building

- Building model using Logistic Regression with RFE
- RFE selected top 15 columns

## 6.Model Evaluation

- Plotted ROC with Gini 0.85, which is a descent value.
- Estimated the Accuracy, Sensitivity and Specificity of the final model on the training set at the optimal cut off 0.35
- Calculated the Precision and recall

- Overall accuracy on train set: 0.7810781078107811
- sensitivity of train set model: 0.7960768287699224
- specificity of train set model: 0.77170582226762
- Precision Score of train set model: 0.7535545023696683
- Recall Score of train set model: 0.6497752349816102

## 7.Predictions on Test Data

- Overall accuracy on Test set: 0.7924459112577924
- sensitivity of our logistic regression model: 0.8120229007633588
- specificity of our logistic regression model: 0.7802263251935676

Sensitivity of our logistic regression model is 81.20%

## 8.Insights from the final model

Variables in model, that contribute towards lead conversion are:

- Affecting Positively:
  - time_on_website
  - last_activity_SMS_Send
  - lead_source_reference
  - lead_source_google
- Affecting Negatively:
  - Last_notable_Activity_modified
  - Olark_chat_conversation
  - do_not_email

## 9.Conclusions and Business Recommendations

### Hot Leads:

The most potential leads are the ones having Predicted Probability > 0.35.

- Use Email to Communicate with the Hot Leads
- Conversion rate increases with increase in the time spend on the website, therefore increase the user engagement in their wesite.
- Try to give SMS notifications,it improves the conversion rate
- Improve the digital marketing to reach out to more people