

# Lead Score Case Study Summary Report

-Arpitha A S

# 1. Problem Statement

- An education company named X Education sells online courses to industry professionals.
  - The company markets its courses on several websites and search engines like Google.
  - When these people fill up a form providing their email address or phone number, they are classified to be a lead
  - 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%.
  - 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 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%.

## Objective :

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

## 2. Solution Approach

### Data Cleaning and Preparation :

- Dropped the features which are part of Lead profiling activities like Tags, Lead Profile, Lead Quality, all Asymmetrique scores.
- Dropped all highly skewed variables which are having more than 90% of the entries related to one category
- Page Views Per Visit, Total Visits & Total Time Spent On Website are the only continues variables retained after dropping unnecessary columns.
- In all the below features, there are certain categories which are very less in count. All those are grouped into single category 'Others'.
  - Last Activity
  - Lead Source
  - Lead
  - Specialization
  - City
- In some of the features there were 'Select' category which are converted into Null values

Below are the Features that will be used for Modelling. All the categorical variables are converted to dummy variables

```
Index(['Lead Number', 'Lead Origin', 'Lead Source', 'Converted', 'TotalVisits',  
      'Total Time Spent on Website', 'Page Views Per Visit', 'Last Activity',  
      'Specialization', 'What is your current occupation', 'City',  
      'A free copy of Mastering The Interview'],  
      dtype='object')
```

- The Data imbalance ratio is about 38% on the converted columns

### 3. EDA :

- With bivariate analysis of categorical variables, the following observations are evident :
  - Lead origin of 'Lead Add Form' influences more for conversion of the lead
  - Welingak Website' and 'References' lead sources are reliable for lead conversion
  - Reference is reliable for a lead conversion
  - SMS Sent' activity is an higher reliable for a lead to be converted
  - Finance Management' specialization is not leading for a good conversion rate. Working professional are reliable for lead conversion
- When the Bivariate analysis performed on the continuous features, its observed that the Total Time Spent on Website influences the lead conversion

### 4. Model Building

- To begin with , Identified 15 variables using RFE method
- After about 5 iterations of eliminating variables with high P-value and high VIF, a final model is achieved with 11 features and listed below :

	coef
-----	-----
const	-2.7953
Total Time Spent on Website	1.1251
Lead Origin_Lead Add Form	3.9313
Lead Origin_Lead Import	1.1811
Lead Source_Olark Chat	1.1451
Lead Source_Welingak Website	1.7279
Last Activity_Email Link Clicked	0.9473
Last Activity_Email Opened	1.5246
Last Activity_Others	1.7926
Last Activity_Page Visited on Website	0.9091
Last Activity_SMS Sent	2.5602
What is your current occupation_Working Professional	2.8651
=====	=====

## 5. Identify the Probability Cutoff

- With ROC area 0.87, the model seems to be good one.
- The Precision\_Recall\_Curve and Sensitivity\_Specificity tradeoff are plotted to identify the optimal cut-off probability so as to keep the metric score good.
- 0.40 is found to be optimal cut-off probability from Precision\_Recall\_Curve and 0.33 from Sensitivity\_Specificity tradeoff
- Objective of the case study is to identify about 80% potential lead candidates. Hence we should target high Recall and should be around 80%. So we stick to probability cut-off of 0.33

## 6. Predications & Metrics on Test Data

- The final model is applied on the test data with the probability cut-off 0.33 and the metric seems to good with not much drop compared to train data.
- Probability also converted to lead score ranging from, 0 - 100 for all the leads of train and test data

## 7. Final Metrics

**Train Set : Recall -79.8% , Accuracy – 80.1%, Sensitivity – 79.8% , Precision – 71.2%, Specificity - 80.2%**

**Test Set : Recall- 80.5%, Accuracy - 80.1% , Sensitivity – 80.5% , Precision – 71.2%, Specificity – 79.9%**

•