# **Lead scoring case study summary**

## **Submitted by:**

**Jyoti**

**Simran**

The provided **‘Leads’** dataset was examined and variables were analysed from provided **‘Leads Data Dictionary’**. Firstly, the **problem statement** given was X education sells online courses. Interested professionals landing on the website are generated as leads. Current conversion rate is 30% and company now want these potential leads to convert into hot leads.

**Business Objective:**

**CEO** wants to turn those 30% into 80% sure shot leads. Wants model to be able to handle future constraints as well as Peak time actions required, to fully utilize man power.

**Step by step approach towards solution:**

Initial process is to **Clean Data,** starting with hot leads that have higher conversion rate resulting in effective use of time. Time spent on nurturing hot leads can be increased than the time spent on cold leads which can be done by building a logistic regression model.

1. **Data Analysis:** columns with higher missing values were present in the dataset. Columns with default value ‘**Select’** is populated so it was filled an ‘**NAN’.** Missing values were imputed with suitable values (median and mode for numeric and categorical respectively) or were dropped if null values were greater than 40%.
2. **Data preparation:** **univariate and multivariate** analysis were carried out.

* Lead conversion rate was 38.54%.
* Country being India and city being Mumbai had the highest conversion rate.
* Boxplots and descriptive statistics indicate outliers present which being removed can lead to data loss by 9%.
* ‘**Finance management’** as **a ‘specialization’** and ‘**Google**’ as ‘**Lead source**’ and **‘unemployed’** as the **‘Occupation’** has highest leads and lead conversion.
* While plotting correlation for numeric variables conversion was correlated with **total website visits, total time spent on website, and inversely correlated with page views per visit. VIF** will be used in further modelling.
* As logistic regression is based on numerical data, converted categorical columns using dummy variables and label encoding and train and test split (70% and 30% resp.).
* Columns with no variance were dropped.

1. **Model Building:** the model was built using RFE and manual fine tuning using P-values and VIF’s. split data into train and test set and make further predictions. Final predictions were made on the final train set and Lead Score were assigned. The accuracy is to be found around 89%. After creating a confusion matrix, sensitivity (86%), specificity (91%), precision (86%), and recall (86%) were calculated. After ROC curve and precision and recall call trade-off, 0.3 was selected as optimal cutoff.
2. **Recommendations:**

* Leads with high ‘lead score’ can be more focused on.
* Marketing from Google has the highest conversion rates as the traffic volume is high
* Encouraging referrals with exciting incentives
* Mumbai has the major leads, can focus on marketing to other cities to achieve higher results.
* Unemployed category has the focus, and finance with specialization
* Students having the least rate.