# Lead Scoring Summary Report

**By**
**Gloriya Thomas**
**Balasubramanian Venkatesan**

This case study aims to help X Education to find the best strategy to explore different ways to encourage industry professionals to join their courses.

Dataset provided gives a lot of information like how often the customers visit X Education web site, how much time they spent on the website etc

***Steps followed in this problem solution:***

1. **Importing Libraries and Inspecting the dataset:**

    - Imported all the necessary libraries and loaded the dataset.
    - Basic inspection of the dataset – 9240 rows and 37 columns.
    - Descriptive statistics of numeric variables - data type and missing values.
    - Total 17 out of the 37 variables with missing values.
    - 7 numeric variables and 30 categorical variables.

2. **Data Cleaning - Missing value & Outlier Treatment**

    - All redundant columns, duplicated rows and unique valued columns were removed.
    - Missing values for numeric variables were imputed with median and categorical variables with mode.
    - Outliers were treated appropriately.
    - 98.5 % rows of the original dataset were retained after imputing all the missing values.

3. **EDA:**

    - Analyzed the dataset with univariate, bivariate and multivariate analysis for both numeric and categorical variables.
    - Multi-variate analysis helped analyze the correlation among numeric variables with the help of pair plot and heatmaps.

4. **Data Transformation / Dummy Variable Creation:**
    - Categorical variables with Yes /No values were converted to 1 and 0 respectively and Dummy variables were created for the remaining categorical variables.

5. **Data Preparation -Train-Test Split and Scaling:**

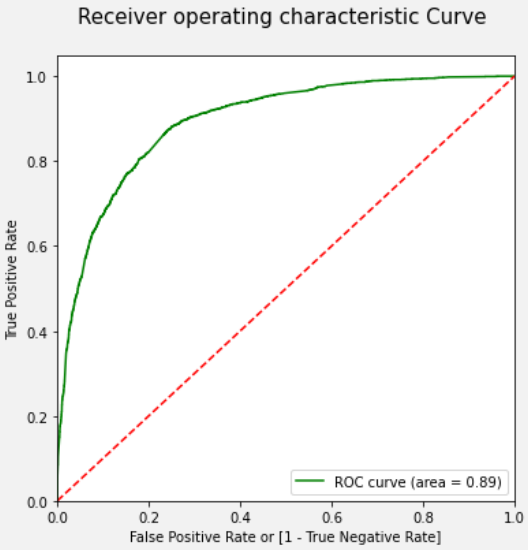
    - The train-test split was done at 70% - 30%
    - Numeric variables were scaled with MinMaxScaler.

6. **Model Building and Final Prediction:**

    - Started model building with the help of RFE to filter the top 15 significant variables.
    - Built the model using the ().
    - Further did an iterative model building process using GLM with repeated feature elimination. Used the p-values and VIF values for the feature elimination.
    - Final model had 12 most significant variables, without losing much accuracy.
    - Checked the precision and recall with accuracy, sensitivity and specificity for the final model and selected an optimum threshold of 0.34 for final prediction.

7. **Model Evaluation:**

    - We created the confusion matrix, draw the ROC curve
    - AOC is .89 indicates the model got a good predictive power.




- Train dataset scores:

```
SCORES FOR THE TRAINING DATA SET

Accuracy   :  80.98 %

Sensitivity :  81.67 %

Specificity :  80.56 %

Precision :  72.23 %

Recall :  81.67 %

F1 score :  76.66 %
```

## 8. Prediction:

- In the final prediction we added a Lead Score for each lead in our dataset, which is an intuitive way to assess an individual's chance of conversion.
- Test dataset scores:

```
SCORES FOR THE VALIDATION DATATSET

Accuracy  :  80.5 %

Sensitivity :  80.78 %

Specificity :  80.33 %

Precision :  71.2 %

Recall :  80.78 %

F1 score :  75.69 %
```

9. **Conclusion**:

As per the model, the most significant features deciding conversions are:

- Total Time Spent on Website
- Lead Origin_lead add form
- Last Activity_had a phone conversation
- What is your current occupation_working professional
- TotalVisits

X Education can rely on our model to have a more efficient approach to target the potential candidates to achieve higher conversion rates and to get more people enrolling to their programs in the future.