# Lead Scoring Case Study Summary Report

This analysis is done for X Education and to find ways to get more customers to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend on the website, how they reached the site, etc., and whether they were converted or not.

Steps Involved in Solving the Case Study:

1. Importing and reading the dataset.
   - Importing the “Leads.csv” dataset into python.
2. Inspecting the dataset.
   - We looked into the shape, info, and statistical aspects of the dataset.
3. Data Preparation.
   - Checked for missing values.
   - Dropped columns having more than 40% missing values.
   - Replacing NaN values with mode/new category values.
   - Checked for outliers and treated them using capping and flooring method.
4. EDA.
   - Created Univariate analysis plots.
   - Created plots with hue = target variable.
   - Created Correlation heatmap and pairplot.
5. Dummy Variable Creation.
   - We mapped Binary variables with two levels (yes & no) to 1’s and 0’s.
   - Created dummy variables for categorical variables.
6. Train-Test Split.
   - We Put all feature variables in X and Target variable in y.

- We Split the dataset into train and test set in the ratio of 70:30.

7. Feature Scaling.
   - Standard scaler scales the features so that the predictors have a mean of 0 and standard deviation of 1.
   - We have done Scaling of numerical features using Standard scaler.
8. Model Building.
   - Using RFE (recursive feature elimination) to select top 15 variables for our model.
   - We built a model with the features selected by RFE using StatsModels.
   - Dropped Insignificant features having high p-value and High VIF value.
   - We repeated the model building process till we arrived to a model with features having normal p-value and VIF values.
   - We then predicted the values on train set.
   - Created a confusion matrix and checked the accuracy, sensitivity and specificity of our model.
9. Plotting ROC Curve.
   - We then plotted a ROC curve, higher the area under the ROC curve the better is your model. The value of ROC curve should be closer to 1, we have the area under ROC curve = 0.96.
10. Finding Optimal Cutoff Point.
   - From the plot we decided to take the cutoff point of 0.2.
   - After taking the cutoff point of 0.2 we created a confusion matrix and checked the accuracy, sensitivity and specificity.

- The accuracy, sensitivity and specificity for train set are 90%, 89%, 91% respectively.
- We also checked for precision and recall scores.

11. Making Predications on the Test set.

- We scaled the variables in the test set.
- We predicted the values on the test set.
- We created a confusion matrix for test set and checked the accuracy, sensitivity and specificity.
- Accuracy, sensitivity and specificity for test set are 89%, 87%, 90% respectively.

Final Results:

**Final Comparison of Observation values between Train and Test dataset.**

**Train Dataset.**

- Accuracy : 90%
- Sensitivity : 89%
- Specificity : 91%

**Test Dataset.**

- Accuracy : 89%
- Sensitivity : 87%
- Specificity : 90%

Submitted By:

Shyam Dalsaniya

Iranna Chatti